International Journal of Computer (IJC)
ISSN 2307-4523 (Print & Online)

http://ijcjournal.org/

Holistic Approach to Big Data Definition using Analysis of Facts

Akande Oyebola[a*], Osofisan Adenike[b], Adekola Olubukola[c]

[a,c]*Computer Science Department, Babcock University, Ilishan Remo, Ogun State, Nigeria*
[b]*Computer Science Department, University of Ibaban, Ibadan, Oyo State, Nigeria*
[a]*Email: akandeo@babcock.edu.ng*
[b]*Email: ao.osofisan@mail.ui.edu.ng*
[c]*Email: adekolao@babcock.edu.ng*

Abstract

Big data has become a concern of science, industry, business, and academics, thus it is no more a buzzword but an emerging technology as viewed by researchers from different perspectives. Thus, different perspectives produced different definitions, of which none of them fully described big data. This research analysed some profound definitions based on discovered facts about big data. The facts are its characteristics, its technology, mode of transfer, its analysis, its infrastructure and security. Thus, a new definition was proposed which captures the basic facts about big data. Big data has its characteristics as foundations, and the rest facts as the pillars. These facts reflect in-depth meaning and understanding of big data to science, industry, business and academics.

Keywords: Analysis; Big data; Infrastructure; Security; Technology.

1. Introduction

The term 'Big Data' first appeared in 1998 in a Silicon Graphics slide deck titled "Big Data and the Next Wave of Infra Stress" by John Mashey [1]. The first book on big data was a data mining book by Weiss and Indrukya in 1998 [2]. As at 2014, nine out of 10 business leaders considered data to be the fourth factor of production, as fundamental to business as land, labour and capital [3]. Big data is the concern of science (for example analysis of human genome, large synoptic survey telescope, network sensors among others.), industry (for example IBM, Google, Microsoft, Facebook, and Telecoms among others.), business (for examples large retail stores such as Shoprite, Walmart among others) and academics.

---

* Corresponding author.

2. Statement of Problem

Big data somewhat described as difficult data is a pertinent concern of science, industry, business, and academics. Hence, different perspectives produced different definitions, such that none seemed to comprehensively describe the subject. Thus, there is a need for a new definition which would capture basic characteristic facts (that is, its characteristics, technology, mode of transfer, analysis, infrastructure and security) about big data.

3. Methodology

These studies embody case studies, systematic literature reviews and surveys. Important requirements were identified in related papers. The relevant documents obtained were qualitatively analyzed for convergence, and relevant details were extracted using inductive approach. No difficulty of being an ethical researcher was encountered.

4. Characteristics of Big Data

Attributes or characteristics of big data are put together as the V's of big data. Reference [4] pronounced the first 3V's, and they are the first three discussed below (volume, velocity and variety).

(a) Volume: This is about the size of data, for examples huge terabytes of data generated in astronomy from telescopes, spectra and so on. Data generated from e-commerce by making the business available to millions of consumers at the same time led to large volume of data in business owners' database. Some researchers measured volume in bytes, while some measured it by number of records, tables or transactions in millions, billions, and trillions. Mostly Researchers use volume to indicate that a dataset is big, but [5], stated that big is a relative word, difficult data is most appropriate to use. Nevertheless, the volume of data is an important attribute of big data.

(b) Velocity: This is the measure of latency or speed at which data arrives, and the challenge of extracting useful information from them at real time. It refers to frequency of data generation or frequency of data delivery [6].Velocity implies both the rate at which data arrive and the time in which it must be acted upon. That is, how fast it arrives and how fast it is consumed.

(c) Variety: This implies that data is from diversity of sources such as sensor, audio, video, graph etc., diversity of formats, quality, and structures. For examples variety of data are generated in ocean science through satellites, gliders etc., and heterogeneity of data types, data representation, and semantic interpretation. Data can be structured (data stored in database with semantic meaning), semi-structure (XML and RSS feeds), or unstructured (for example: texts, calls, tweets, transaction made using credit card or debit card) data with no latent meaning [1]. Data from diverse sources for a particular task may have the challenge of integration of these data. Data in different format have the challenge of best normalization method to use.

Other emerging 3 Vs ([7;1]) are:

(d) Veracity or Variability: This implies changes in the structure of data and its interpretation [8]. That is, certainty / uncertainty of data, validity / invalidity of data, accuracy / inaccuracy of data, in terms of quality, relevance, predictive value and meaning [9].

(e) Value: This implies the value placed on data by various stakeholders such as Government, Industries such as bank and manufacturing, Health Care such as hospitals, pharmacy etc [9].

(f) Valence: This implies connectedness or relationship among data, for example, connectedness or relationship of people on social media such as Facebook, LinkedIn, and Research Gate. The connectedness can be through social status, school, age, and workplace [1]. Relationship or connectedness can be direct or indirect.

In reality, the combinations of all the characteristics may not be feasible, thus Figure 1 is an illustration of the 6 V'S showing their various combinations. BD indicates combinations of characteristics that a dataset can have to be called big dataset. Note that volume alone is sufficient as a characteristic for a dataset to be called big. Wherever the inward textured circle reflects on the combination of characteristics represents big data, and the arrow points to where no combination of characteristics lies on the textured circle, thus no big data. However, the textured circle reflects on variety, despite variety alone cannot make a data to be called big.

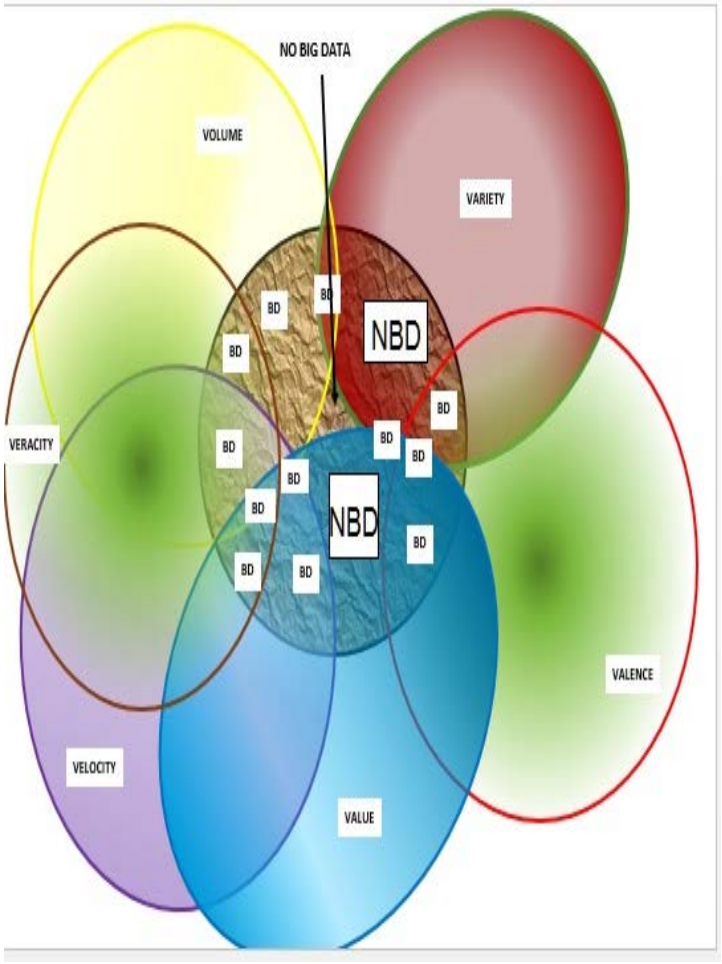

Figure 1: Illustration of Big Data Characteristics

5. Big Data Definitions

Researchers viewed big data from various perspectives and, therefore, described it as data intensive technology [10]. Big data was viewed as an environment, an architecture, infrastructure (i.e. data center) [11], analysis (i.e. data science), processes and processing speed, growing data scale [12], data management, from its challenges, its sources and acquisition, complexity [13], and as structure or unstructured datasets [12] among others. Thus, its definitions are in different forms.

Reference [12], stated that "Big data refers to huge data sets that are orders of magnitude larger (volume); more diverse, including structured, semi-structured, and unstructured data (variety); and arriving faster (velocity) than you or your organization has had to deal with before". Reference [12] only dealt with three characteristics: volume, variety, and velocity. It did not include value, and veracity, in addition, the definition did not include the technological features, and method of analysis.

According to [11], big data referred to different types of data: traditional enterprise data, machine-generated/sensor data, and social data. Reference [11] was about sources of big data. Reference [14] stated that "Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value". [14] was not specific about type of technology it uses, and the source of big data. Reference [15] defined that "Big data is a term describing the storage and analysis of large and or complex data sets using a series of techniques including, but not limited to: NoSQL, MapReduce and machine learning". Reference [15] did not include the source, value or impact of big data analysis on society, organization, or business process.

Reference [5], defined big data as any data that is expensive to manage and hard to extract from. Reference [3] defined that "big data as not only referring to the data itself, but includes the set of technologies that can capture, store, manage and analyse varied collections that solve complex problems and make unlocking value from that data more economical". Reference [3] defined big data as a term to describe large datasets that could not be captured, stored, managed nor analysed using traditional databases. Thus, the authors did not include the characteristics of big data, except volume.

From the extant research perspective, big data encompasses both or either data intensive and computational intensive technologies whose analysis is the concern of machine learning and artificial intelligence techniques (using both or either machine learning or artificial intelligence techniques for analysis) in order to discover hidden knowledge. Thus, this current research proposed a definition as:

Big data technology involves ***growing scale of data*** *or* ***complexities of data*** *which may or may not be* ***connected****, but can be* ***stored*** *and* ***analysed*** *either in* ***real time or in batches*** *through* ***data intensive and/or computational intensive technologies*** *using* ***machine learning or artificial intelligence techniques*** *in order to* ***acquire hidden unknown knowledge*** *from data without neglecting its* ***security****.*

Table 1 is an analysis of the proposed definition of big data from this current research.

Table 1: Analysis of Big Data definition

KEYWORDS / TERM	***SIGNIFICANCE***
growing scale of data	*Volume*
complexities of data	*Variety which includes*
Connected	*Valence, that is*
stored	*Storage medium which*
analysed	*Big data analytics*
real time or in batches	*Velocity of data*
data intensive and/or	*Technology for storage,*
machine learning or	*Techniques for data*
acquire hidden	*Value or impact of big*
security	*Big data security and*

Figure 2 depicts characteristics of Big data has its foundations and other facts as the pillars:

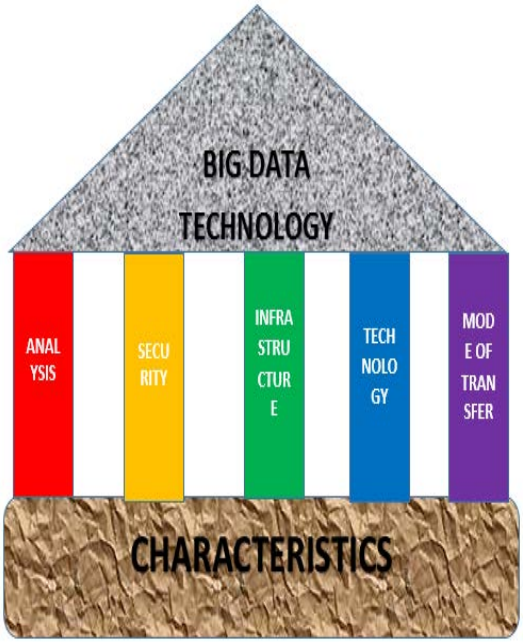

Figure 2: The Foundation and Pillars of Big data

5.1 Importance of big data

a) It provides possibility of discovering subtle patterns that are not possible with small-scale data [16].
b) It helps enterprise to have a more insightful understanding of their business in order to have better innovation, enhanced productivity, and stronger competitive power [11].

c) In manufacturing industry, big data reveals usage pattern, failure rates, and gives insight into product improvement [11].

6. Big Data Challenges

Big data is not without its challenges that emanates from data analytics, data storage (system capabilities in terms of storage and memory), data security, algorithm design, business model ([12]; [17]), and its characteristics [1]. The followings are some of the challenges of big data discovered by this current research:

(a) Data Storage: This challenge emanates from volume -the first characteristic of big data. As the volume of data increases, so the challenge of where to store them efficiently increases. Attempt to provide solution to this lead to data intensive technologies and cloud computing.

(b) Data Latency: This concerns the speed of data arrival and retrieval. This challenge hangs on two major characteristics; volume, and velocity. It has to do with system capability in terms of memory.

(c) Data Security and Privacy: This implies security of storage facility, and privacy of individual or company or country or national data. It includes security of data infrastructure (that is data center) such as infrastructures in the cloud, and reliability of data science tools and platforms, such as Hadoop.

(d) Data Analytics: The challenge of big data analytic emerged from all the characteristics of big data such as volume, velocity, variety, veracity/variability, valence and value. Volume of big data brings the challenge of intensive computation of data during analysis of big data using machine-learning algorithm. The following section discusses big data analytics because its challenges hang on all the characteristics of big data.

7. Big Data Analytics

Generally, data analytics is the science of examining raw data with the purpose of extracting useful information or knowledge to draw conclusions or for making decision. There are three types of data analytics functions: to make predictions, classifications, and exploratory study [18].

Big data analytics refers to determining, assessing, and interpreting meaning from data, of which the interpretation can be predictive, descriptive, or prescriptive [19]. Several tools from different techniques such as data mining, statistical analysis, artificial intelligence, SQL database, NoSQL database, and parallel processing framework such as MapReduce framework can be employed in big data analytics [6]. Big data analytics is a combination of advanced analytic techniques and big data, thus, big data analytics can be defined as operation of advanced analytic techniques on big data [6].

Basic tools for data analysis are; data mining techniques, machine learning techniques, artificial intelligence, query and reporting, data visualization tools such as excel, PowerBI by Microsoft, Tableau, natural language processing etc.

8. Conclusion

Handling of Big data is currently of special interest to science, industry, business, and academics. Thus, this paper provides an insightful definition based on its characteristics (6 V'S), its technology, mode of transfer, its analysis, its infrastructure and security. Therefore, the definition and identified challenges can provide area of focus to researcher and other interest groups.

9. Recommendation and Further Work

This research has been able to identify different areas of big data challenges namely Data Storage, Data Latency, Data Security and Privacy, and Data analytics. This will aid researchers to focus on specific areas of interest. But the limitation here is that the first three were not elaborately discussed. This research is currently working on challenges emerging from big data analytics based on big data characteristics and technology. In addition, researchers would find the modelled foundation and pillars of Big data in this study as a useful inspiring construct.

References

[1] G. Kaustav and N. Asoke. "Big Data : Security Issues, and challenges." International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), 2016, pp. 1-11.

[2] Y. Prasad and K. R. Reddy. "Data Mining With Big Data." International Journal and Magazine of Engineering, Technology, Management and Research, 2015, pp. 493-496.

[3] F. Helena, F. Evelyn, R. Donya, and B. Dmitriy. Big Data How it can become a differentiator. USA: Deutsche Bank AG, 2014.

[4] L. Dough.. 3D Data Management: Controlling Data Volume, Velocity, and Variety . Stanford: META Group, 2001.

[5] B. Howe. "Big Data Science Needs Big Data Middleware." Seventh Biennial Conference on Innovative Data Systems, 2015.

[6] H. Philipp and M. Heather. "Parallelizing Machine Learning– Functionally: A Framework and Abstractions for Parallel Graph Processing." 2nd Annual Scala Workshop, 2011.

[7] C. Min, M. Shiwen, and L. Yunhao. "Big Data:A Survey." Mobile Netw Appl, 2014, pp.171-209.

[8] F. Wei and B. Albert. "Mining Big Data: Current Status, and Forecast to the Future." SIGKDD Explorations, 2013, pp. 1-5.

[9] R. Hermon, and P. A. William. "Big Data In Healthcare: What Is It Used For." Australian eHealth Informatics and Security Conference. Perth, Western Australia: Edith Cowan University Research

Online, 2014, pp.39-49.

[10] L. Jimmy and ChrisDyer. Data-Intensive Text Processing with MapReduce. Morgan and Claypool, 2010.

[11] J. Dijcks. Oracle: Big Data for the Enterprise. Carlifonia: Oracle Corporation, 2013.

[12] Intel. 2014. Peer Research: Big Data Analytics. USA: Intel, 2014.

[13] R. Andreas. (2016, November 2). "Big Data Definition." MIKE2.0. November 2. Available: http:mike2.openmethodology.org/wiki/Big_Data_Definition. [July 1, 2018].

[14] D. M. Andrea, G. Marco, and G. Michele. "What is Big Data? A consensual Definition and a Review of Key Research Topics." International Conference on Integrated Information, AIP, 2014, pp. 97-104.

[15] J. Ward and A. Baker. "Undefined By Data: ASurvey of Big Data Definitions." Cornel University Library. September 20. arXiv:1309.5821v1, 2013.

[16] F. Jianqing, H. Fang and L. Han. "Challenges of Big Data Analysis." arXiv, 2013, pp. 1-38.

[17] Morais, Telmo da Silva. "Survey on Frameworks for distributed Computing: Hadoop, Spark and Storm." 10th Doctoral Symposium on Informatics Engineering. Porto, Portugal, 2015, pp. 95-105.

[18] A. Bansal, M. Shama, and S. Goel. “Improved K-mean Clustering Algorithm for Prediction Analysis using Classification Technique in Data Mining.” International Journal of Computer Applications, 2017, 35-40.

[19] A. Manohar, P. Gupta, V. Priyanka and M. Uddin. Utilizing Big Data Analytics to Improve Education. Bridgeport, Bridgeport, CT USA, 2017.