ISSN: 2454-1907
DOI: https://doi.org/10.29121/ijetmr.v5.i9.2018.287

IJETMR

International Journal of Engineering Technologies and Management Research
A Knowledge Repository

BIG DATA ANALYTICS: A PRIMER

Matthew N.O. Sadiku [1], Justin Foreman [*2], Sarhan M. Musa [3]
[*1, 2] Dept of Electrical and Computer Engineering, Prairie View A&M University, U.S.A.
[3] Roy G. Perry College of Engineering, Prairie View A&M University, U.S.A.

Abstract:
The use of digital devices and systems such smart phones, computers, the Internet, and social media has resulted in a massive volume of data which is exponentially increasing daily. Such data is processed using multiple techniques, collectively known as big data analytics. Big data analytics is the process of examining large amounts of data (big data) to uncover hidden patterns, correlations, and other insights. Analyzing big data enables organizations and businesses to make better and faster decisions. This paper briefly presents the fundamental concepts of big data analytics and its tools.

Keywords: Big Data; Big Data Analytics; Advanced Analytics.

Cite This Article: Matthew N.O. Sadiku, Justin Foreman, and Sarhan M. Musa. (2018). "BIG DATA ANALYTICS: A PRIMER." *International Journal of Engineering Technologies and Management Research*, 5(9), 44-49. DOI: https://doi.org/10.29121/ijetmr.v5.i9.2018.287.

1. Introduction

Data is everywhere and it provides information. The volume of data has exploded to unimaginable levels in the past decade. Big data comes from different sources such as sensors, devices, computer networks, machines, IoT, GPS, RFID, ecommerce transactions, web, weather data, medical data, insurance records, and social media. As we capture terabytes of data, our major challenge is making sense of this gigantic amount of data. This is where big data analytics fits into the picture.

Big data analytics deals with examining massive amounts of data- big data- to uncover hidden patterns, market trends, customer preferences, and other useful information that can help in making informed decisions.

2. Big Data

Some believe that the concept of big data originated from Internet corporations such as *Amazon, Google, Netflix, Yahoo,* and others. Big data is a problem in search of a solution. By definition, big data is data that is too large for traditional analysis methods and techniques. It may be structured, unstructured, or semi-structured. Sources of big data can be either collective gathering or individual generation.

ISSN: 2454-1907
DOI: 10.5281/zenodo.1443531

As shown in Fig. 1[1], big data can be characterized as having 5Vs: volume, velocity, variety, value, and veracity.

- *High volume*—the amount or quantity of data
- *High velocity*—the rate at which data is created
- *High variety*—the different types of data (structured, unstructured, or semi-structured)
- *High* Value – the added value from the information extracted.
- *High* Veracity – the uncertainty, accuracy, and reliability of data

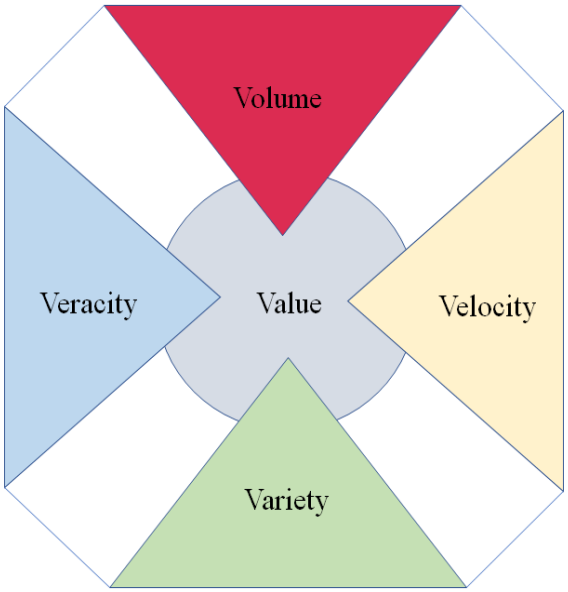

Figure 1: The 5 V's of Big Data [1]

Because of these characteristics, big data exceeds the processing capacity of conventional database systems. It requires new technologies and techniques to capture, store, and analyze. This is where big data analytics techniques and tools come into the picture. Analytics tools and databases can handle big data.

3. BDA Techniques

Big data analytics (BDA) refers to how we can extract, validate, translate, and utilize big data as a new currency of information transactions. It is an emerging field that is aimed at creating empirical predictions. Data-driven organizations use analytics to guide decisions at all levels.

Common BDA techniques include [2,3]:

1) **Data Mining:** Data mining is searching for useful information in massive amounts of data. It involves extracting knowledge from large databases and applying the knowledge in making fact-based decisions. Data mining techniques enable humans to find useful hidden trends and relationships in massive data. Data mining algorithms allow us to refine data and find their real value. They deal with a large amount of data with great processing speed.

ISSN: 2454-1907
DOI: 10.5281/zenodo.1443531

Data mining can be predictive or descriptive. Descriptive data mining describes the general characteristics of the data in the database, while the predictive data mining uses the current data to make some prediction. Data mining techniques include cluster analysis, association rule of learning classification, and regression [4,5].

2) **Web Mining:** This is the technique for extracting useful information obtained from the World Wide Web. It is the application of data mining techniques for discovering patterns from large web repositories. It may be used to learn about customer behavior, evaluate the effectiveness of a specific website, help quantify the success of a marketing strategy, and reveal unknown knowledge about a particular website. Web mining may include content mining, hyperlink structure mining, and usage mining [6,7].
3) **Text Analytics:** It is generally accepted that structured data represents only 20% of the information available to an organization, while 80% of all the data is in unstructured form. Also known as text mining or natural language processing, text analytics is the science of turning unstructured text into structured data. Its purpose is deriving high-quality structured data from unstructured (textual) information.
4) **Predictive Analytics:** This is the process of predicting phenomena of the future with BDA tools by analyzing current and historical facts. This makes it possible for analysts to make sound decisions based on visualization analysis and data mining. It builds models for forecasting customer behavior and other future developments.
5) **Machine Learning:** This is one of the main drivers of the BD revolution because of its ability to learn from data and provide decisions, insights, and trends. Machine learning (ML) is about learning some properties of a data set and applying them to new data. It is the scientific discipline dealing with the ways in which machines learn from experience. It refers to the automated detection of meaningful patterns in a given collection of data. It is an incredibly powerful tool that is used in a wide range of compelling application domains because it is applicable to many real-life problems. ML has already shown its capacity to learn and analyze, beating our best champions at complex strategy games, such as *Go* and *Jeopardy!* Common machine learning algorithms include artificial neural networks, fuzzy logic, and genetic algorithms [8,9].
6) **Deep Learning:** Deep learning (DL) refers to a family of approaches that have taken machine learning to a new level, helping computers make sense out of vast amounts of data. Deep learning algorithms are used to train deep networks with large amounts of data. DL has become a big wave of technology trend for big data and artificial intelligence [10].
7) **Visual Analytics:** Data visualization plays a major role in understanding and exploring data because there is much to gain when data is presented in a visual manner. Visual analytics are efficient when working in a geospatial domain and multi-dimensional analysis.
8) **Crowdsourcing:** This is the process of getting work done by online community or crowd of people in the form of an open call, the voluntary undertaking of a task. This tool is used more for collecting data than for analyzing it.
9) **Mobile Analytics:** Mobile analytics research is emerging in different areas such as mobile sensing apps that are location-aware and activity-sensitive. The ability to collect fine grained, location-specific, context-aware, highly personalized content through these smart devices has opened new opportunities [11].

ISSN: 2454-1907
DOI: 10.5281/zenodo.1443531

Analysing huge volumes of data improves data-driven decision-making mechanism and the ability to predict future outcomes. Big data analytics helps organizations in significant cost reduction, better decision making, and creation of new products and services.

4. BDA Tools

The 5V characteristics of big data will bring software based on traditional approaches to their knees. Once the big data is ready for analysis, we use advanced software programs. These include Hadoop, MapReduce, MongoDB, and NoSQL databases.

- **Hadoop:** This is an open-source software-framework for distributed storage of large datasets on computer clusters. It was originally developed in 2006 by Doug Cutting and Mike Cafarella. It provides large amounts of storage for all sorts of data along with the ability to handle virtually limitless concurrent tasks. Hadoop is written in Java and provides Java classes and APIs to access them. It is designed to process large amounts of structured and unstructured data. Although Hadoop is a free and an open-source software, the free version of Hadoop is not easy to use. A number of companies have developed friendlier versions of Hadoop, and Cloudera is the most popular of them all.
- **MapReduce:** This works on Hadoop framework. It is a Google technology for processing massive amounts of data. It is a software framework that enables developers to code programs that can process large amounts of unstructured data. It is a data processing algorithm that involves distributing a task across multiple nodes running a "map" function. It has two components: (1) Map which distributes the input data to several clusters for parallel processing, (2) Reduce which collects all sub-results to provide the final result.
- **NoSQL:** This is an open source software designed for use in big data application in clustered environments, providing high speed access unstructured or semi-structured data. It provides capabilities to query and retrieve unstructured and semi-structured data. It is available in both an open source community edition and in a priced enterprise edition.
- **MongoDB:** This is a good resource for managing data that is frequently changing or unstructured. It is a version of NoSQL. It is the modern, start-up approach to relational databases. As with any database, one needs to know how to query it using programming language. It is a flexible, highly scalable database designed for web applications. It is often used to store data in mobile apps, product catalogs, and real-time applications.

Other tools include Hive, Cassandra, Spark, Tableau, and Talend.

5. Applications

Applications of BDA allows data scientists, predictive modelers, statisticians, and other professionals to analyze growing volumes of data. BDA has increasingly been employed by retailers, financial services firms, healthcare organizations, manufacturers, energy companies, and other enterprises.

Marketing: Market leaders gain insight by analyzing transaction histories and web-behavior. They use the result to paint a complete picture of each customer's behavior, likes, and dislikes. This allows them to present personalized, tailored offerings to individual customers.

ISSN: 2454-1907
DOI: 10.5281/zenodo.1443531

Fraud detection: Cell phone operators know customer location and credit-card companies know the location of transactions. By analyzing point of sale, geolocation, authorization, and transaction data, business company can identify fraud patterns in historical data. The result of the analysis can be used to identify potential fraud and proactively notify customers.

Healthcare: The healthcare industry has generated large amounts of data, driven by record keeping, compliance, and patient care. By effectively using BDA, healthcare organizations stand to realize significant potential benefits [12].

Cellular network: Mobile cellular networks generate massive heterogeneous data due to the widespread use of mobile Internet. BDA can improve the performance of mobile cellular networks and maximize the revenue of operators [13].

Other applications include telecommunications, transportation, law enforcement, manufacturing, utilities, revenue generation, gaming, automobile insurance, smart grid, and gaming.

6. Benefits and Challenges

The major benefits of big data analytics are speed and efficiency. BDA helps organizations harness their data and use it to identify new opportunities. This results in smarter business moves, more efficient operations, cost reduction, increased productivity, better decision making, higher profits, and happier customers as illustrated in Figure 2.

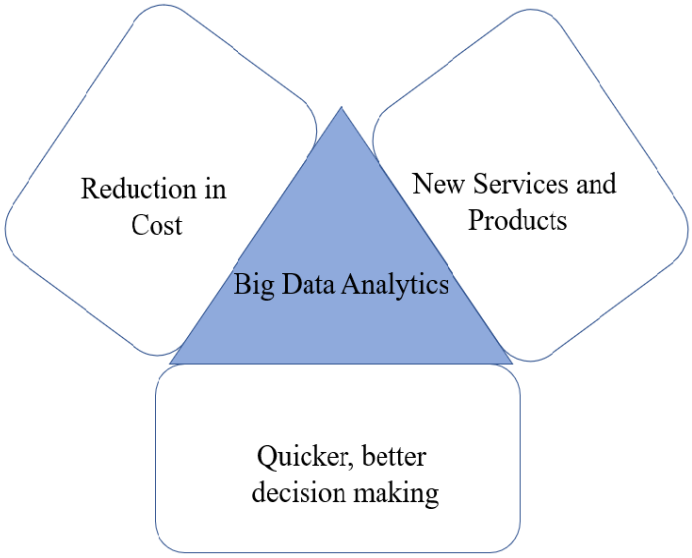

Figure 2: Benefits of BDA [14]

The major organizational challenges to BDA adoption include [15]: (1) Data ownership and control, (2) skills shortage, (3) business focus and prioritization; (4) privacy. Most organizations lack employees who are skilled in big data analytics and find hiring experienced data scientists to be expensive. The amount and complexity of data can cause data management issues such as data quality, consistency, and governance. Individual privacy is a major challenge since a lot of big data contains personal information about customers, clients, or patients. How much data should the government and organizations be allowed to collect? Decision makers must consider these concerns by implementing protection mechanisms that enable getting benefits from big data without risking security and privacy [16].

ISSN: 2454-1907
DOI: 10.5281/zenodo.1443531

7. Conclusion

We live in the big data era. Big data consists of huge amounts of structured and unstructured data. BDA tools provide a means of analyzing such data and drawing conclusions about them to help organizations make informed decisions. These powerful tools will be useful in various industries. They have been revolutionizing various aspects of innovation, research, development, and management. Organizations and businesses have started to invest heavily in big data initiatives. They are gaining insights into customers and operations because of the ability to analyze their massive data. For this reason, more forward-looking organizations are seeking for data scientists who can make sense of their huge data.

References

[1] Song, I.Y & Zhu, Y. 2016. Big data and data science: what should we teach?" Expert Systems, 33, no. 4, August 2016, pp. 364-373.
[2] Sadiku, M.N., Nelatury, S.K & Musa, S.M. 2017. Wireless Big Data. Journal of Scientific and Engineering Research, vol. 4, no. 9, 107-110.
[3] Yaqoob et al. (2016). Big data: from beginning to future. International Journal of Information Management, 36, 1231-1247.
[4] Sadiku, M.N.O, Shadare, A.E. & Musa, S.M. 2015. Data mining: a brief introduction. European Scientific Journal, 11, no. 21, 509-513.
[5] Sadiku, M.N.O., Musa, S.M. & Musa, O.M. 2017. Data mining in the chemical industry. International Journal of Trend in Research and Development, 4, no. 6, 295-296.
[6] Gupta, R. (2014) Journey from data mining to web mining to big data. International Journal of Computer Trends and Technology, 10, no. 1,18-20.
[7] Mohata,P.B. (2015). Web data mining techniques and implementation for handling big data. International Journal of Computer Science and Mobile Computing, 4, no. 4, 330-334.
[8] Sadiku, M.N.O, Musa, S.M. and Musa, O.S. (2017). Machine Learning. International Research Journal of Advanced Engineering and Science, 2, no. 4, 2017, 79-81.
[9] L'Heureux, A. et al. 2017. Machine learning with big data: challenges and approaches. IEEE Access, 5, 2017, 7776-7797.
[10] Sadiku, M.N.O., Tembely, M. and Musa, S.M. (2017). Deep learning. International Research Journal of Advanced Engineering and Science, 2, no. 1, 77-78.
[11] Chen, H., Chian, R.H.L. & Storey, V.C. 2012. Business intelligence and analytics: from big data to big impact. MIS Quarterly,36, no. 4, 1165-1188.
[12] Raghupathi, W. and Raghupathi, V. (2014). Big data analytics in healthcare: promise and Potential. Health Information Science and Systems, 2, no.3.
[13] He, Y. (2016). Big data analytics in mobile cellular networks. IEEE Access, 4, 1985-1996.
[14] SAS. (2017). Big data analytics: What it is and why it matters. https://www.sas.com/en_us/insights/analytics/big-data-analytics.html
[15] Malaka, I. and Brown, I. (2015). Challenges to the organizational adoption of big data analytics: A case study in the South African telecommunications industry. Proceedings of the 2015 Annual Research Conference on South African Institute of Computer Scientists and Information Technologists.
[16] Gahi, Y., Guennoun, M. and Mouftah, H.T. (2016). Big Data Analytics: security and Privacy Challenges. Proceeding of IEEE Symposium on Computers and Communication.

*Corresponding author.
E-mail address: jfforeman@ pvamu.edu