*Invention Journal of Research Technology in Engineering & Management (IJRTEM)*
*ISSN: 2455-3689*
*www.ijrtem.com Volume 2 Issue 10 ǁ October 2018 ǁ PP 93-95*

# Big Data Analytics

[1,]Chandra M. M. Kotteti, [2,]Matthew N. O. Sadiku, and [3,]Sarhan M. Musa
*[1,2,3,]Roy G. Perry College of Engineering Prairie View A&M University Prairie View, TX 77446*

***ABSTRACT:*** *In the present generation, data has become the basis for most of the innovations in advanced sciences and technologies. However, the data that is used for analysis is huge and contains disparate data fields. Analyzing this data is a non-trivial task. Nevertheless, with the advancement in big data research, new technologies have been invented to handle problems related to big data. One such technology is Big Data Analytics. It has a wide application scope and very useful in future prediction tasks using the data in hand. In this work, we briefly introduce Big Data Analytics and its importance as well as some of its applications and challenges.*

***KEY WORDS:*** *data pre-processing, big data analytics, information technology, decision-making systems*

## I. INTRODUCTION

Nowadays, the term *'data'* is well-known in every business, especially among Information Technology (IT) companies. Data collection and analysis has become a common trend and an integral part of every business. In fact, the data collected even by a small-scale company is humongous and is beyond human abilities to analyze. Why do companies collect and analyze data about their business? Because, data acts as the basis for predicting the future of their business growth. However, data collection and analysis in a non-trivial task, it includes three main challenges: storage space, data pre-processing, and computational power. With the help of advanced technologies, more powerful and efficient smart devices are being effortlessly produced and they are made readily available to almost every human-being. For example, a smart phone can produce good amount of data in one single day of its use. Imagine how big the data which is generated by a company in every day of its operation using several computers, servers, etc. Therefore, storage space plays a key role in data collection task. Advancement in semi-conductor industry is supporting companies to save and process tons of data seamlessly. On the other hand, raw data collected is usually expected to be unstructured, noisy, and consisting of a variety of data types. For instance, raw data can have missing values, and contain variety of data types such as string, numeric, image, etc. Data pre-processing helps in cleaning the raw data, making it structured and noise-free, and extracts useable and relevant features from it to better support business decision-making systems. In a real-time business scenario, the speed at which data being produced is swift. In a typical client-server environment, the communication between the client and server should be spontaneous. Client sends a request to server; the server has to be capable enough to respond immediately. Otherwise results that poorly impact business growth may occur. Computational power decides how quickly an end-to-end transaction can happen between a client and a server. It is responsible for rapid data collection, storage, analysis, and for generating quick response.

**Big Data :** Big Data is a combination of several technologies typically deals with large datasets that is used to gain deeper insights into data, to visualize market trends, identify hidden patterns and unknown data correlations, and other important business information to benefit a company's business growth. It involves analyzing huge volumes of data containing several data types, at the right speed and right time to support real-time analysis and response [1].

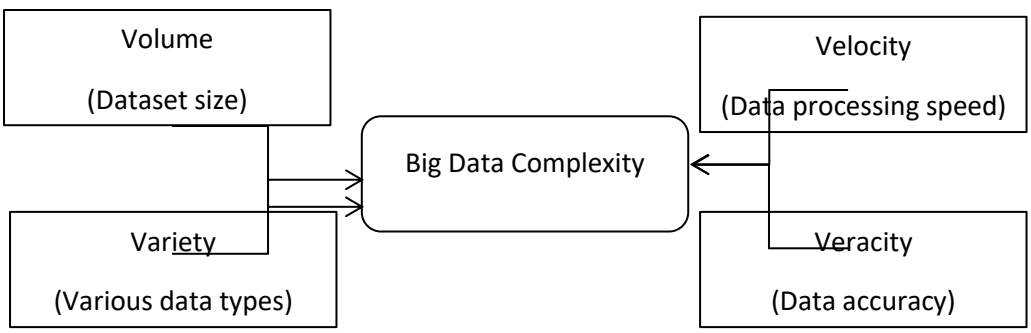


Fig 1. Four V's of Big Data [1]

As shown in Fig 1. the complexity of Big Data can be determined using four V's: volume, variety, velocity, and veracity. Data analysis model training depends mainly on the size of the input data (volume), less amount of data may result in model under-fitting issues, and inputting loads of data into a model may overfits it. Hence, depending on the type of application proper input data size should be selected.

Big data analysis models are expected to have high standard computational power to maintain high data processing speed (velocity). In real-time applications, even a few seconds of delay in response is unacceptable. They should also possess data handling tools and techniques to pre-process disparate data (variety), before it can be used for model training. The quality of the data (veracity) selected for model training decides the model output, irrelevant input data negatively affects model's performance.

**Key things of Big Data Evolution**

According to [2], the key things of Big Data Evolution include:

- Transforming data to wisdom: Information is extracted from collected data, and using that information as basis knowledge is build-up, which ultimately gives wisdom of when to use that knowledge.
- Data storage: In the recent past, Relational Database Management Systems (RDBMS) overcome the redundancy issues in traditional database management systems. However, they fail to support custom data types. To address this limitation Object Relational Database Management System (ORDBMS) has been developed, which supports any custom data types.
- Attributes of data: Important attributes of data include; data cleanliness, data consistency in its representation, data accuracy, and comprehensiveness.
- Data processing: On Line Transaction Processing (OLTP) and On-Line Analytical Processing (OLAP) systems are well-known data processing techniques, where former is used to process several on-line transactions and latter uses complex queries for data mining and analytics.
- Data warehouse: OLTP systems lack power to carry out analytical functions on large datasets. Extract Transform Load (ETL) process has been developed to build data in a Data Warehouse using operational databases to support Decision Support Systems (DSS). OLAP model is applied to Enterprise Data Warehouse (EDW) to analyze enterprise level data.
- Grid Data Computing (GDC): It uses a set of resources and computers to achieve a common goal. Distributed Data Computing (DDC) manages a pool of computers with message passing technique to achieve the required objective. Cloud users in Cloud Data Computing (CDC) use shared resources to make data processing more economical.

**Application scope:** There are many factors influencing the application scope of big data analytics. For example, in complex systems, data-driven modeling and hypothesis generation is important to understand system behavior and interactions as well as to achieve controlled evolution and design [3]. According to [4], organizations use big data analytics to gain knowledge that is hidden inside the data to help in predicting customer's purchasing behavior, fraud detection, understanding business needs, and generating big revenues.

In Banking sector, banks have great advantage of using big data analysis, because of the huge amount of data they have been collecting for decades. It could be helpful to observe money movements, provide protection against thefts, and understand consumer behavior [5]. Moreover, over the years big data analytics has been gaining popularity in agriculture related applications. It helps in giving advance weather predictions, enhancing yield productivity, reducing unnecessary harvesting costs, etc. [6]. Some applications domains need fast execution of data collection and analysis processes. For example, in Finance, large amount of dynamic data is generated rapidly. Companies use this dynamic big data to help in early detection of opportunities and threats for increasing their profits. However, in real time situations, data transfer, situation discovery, analytics, and decision making become challenging for time critical big data applications [7].

**Research challenges**

According to [8], research challenges could be:

- Data privacy, security, and protection: This becomes important, especially when data being analyzed is sensitive, for instance, online bank transactions contain highly sensitive information, which are to be carefully handled to avoid data breach.
- Cleaning raw data: Sometimes raw data collected contains data features which are not significant. Extracting useful data features at the same time filtering out un-useful data is a challenging task, but good data pre-processing help in reducing computational complexity of data analysis models.

- Automatically generating right metadata: To describe what data is collected, how it is stored and measured. It is useful to handle data which is not in a suitable format ready for analysis.
- Data mining: It requires high quality data for analysis, effective data querying interfaces, scalable mining algorithms, and efficient computing environments. Development in any of these requirements for data mining will help in performing better data analysis.

## II. CONCLUSION

Every day, data (digital) is growing bigger and bigger, and big data analysis has become a requirement for gaining invaluable insights into data such that companies and organizations could gain significant profits in the global market. Big Data Analytics is a technology that aims to support businesses to extract knowledge from data, which can be used for future prediction, increasing profits, and enhancing customer service.

## REFERENCES

[1] J. Hurwitz, A. Nugent, F. Halper and M. Kaufman, Big Data For Dummies, Hoboken, NJ: John Wiley & Sons, 2013.

[2] K. Venkatram and M. A. Geetha, "Review on Big Data Analytics - Concepts, Philosophy, Process and Applications," Cybernetics and Information Technologies, vol. 17, pp. 3-27, 2017.

[3] K. Kambatla, G. Kollias, V. Kumar and A. Grama, "Trends in big data analytics," Journal of Parallel and Distributed Computing, vol. 74, pp. 2561-73, 2014.

[4] M. A. Memon, S. Soomro, A. K. Jumani and M. A. Kartio, "Big Data Analytics and Its Applications," CoRR, vol. abs/1710.04135, 2017.

[5] U. Srivastava and S. Gopalkrishnan, "Impact of Big Data Analytics on Banking Sector: Learning for Indian Banks," Procedia Computer Science, vol. 50, pp. 643-52, 2015.

[6] M. R. Bendre, R. C. Thool and V. R. Thool, "Big data in precision agriculture: weather forecasting for future farming," in 2015 1st International Conference on Next Generation Computing Technologies (NGCT). Proceedings, Piscataway, 2015.

[7] N. Mohamed and J. Al-Jaroodi, "Real-time big data analytics: applications and challenges," in 2014 International Conference on High Performance Computing & Simulation (HPCS), Piscataway, 2014.

[8] Y. He, F. R. Yu, N. Zhao, H. Yin, H. Yao and R. C. Qiu, "Big Data Analytics in Mobile Cellular Networks," IEEE Access, vol. 4, pp. 1985-96, 2016.

## ABOUT THE AUTHORS

**Chandra M. M. Kotteti** is currently a doctoral student at Prairie View A&M University, Texas. His research interests include fake news detection using machine learning and deep learning, natural language processing, big data analytics, and wireless networks.

**Matthew N.O. Sadiku** is a professor at Prairie View A&M University, Texas. He is the author of several books and papers. He is an IEEE fellow. His research interests include computational electromagnetics and computer networks.

**Sarhan M. Musa** is a professor in the Department of Engineering Technology at Prairie View A&M University, Texas. He has been the director of Prairie View Networking Academy, Texas, since 2004. He is an LTD Sprint and Boeing Welliver Fellow.