International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST

Big Data and Seven (7) V's Characteristics in Industry

S.Suganya[1*], Dr.T.Meyyappan[2]
[1]Research Scholar, [2]Professor & Head,
Department of Computer Science, Alagappa University, Karaikudi, India
* Corresponding author e-mail: suganyasudhakar04@gmail.com)

Article Info
Volume9, Issue 6
Page Number: 630-636

Publication Issue
November-December-2022

Article History
Accepted : 05 Nov 2022
Published : 30 Dec 2022

ABSTRACT

Big data has become an important issue for a large number of research areas such as data mining, machine learning, computational intelligence, information fusion, the semantic Web, and social networks. The rise of different big data frameworks such as Apache Hadoop and, more recently, Spark, for massive data processing based on the Map Reduce paradigm has allowed for the efficient utilization of data mining methods and ma-chine learning algorithms in different domains. A number of libraries such as Mahout and Spark ML ib have been designed to develop new efficient applications based on machine learning algorithms. The combination of big data technologies and traditional machine learning algorithms has generated new and interesting challenges in other areas as social media and social networks. These new challenges are focused mainly on problems such as data processing, data storage, data representation, and how data can be used for pattern mining, analyzing user behaviors, and visualizing and tracking data, among others. In this paper, we present a revision of the new methodologies that is designed to allow for efficient data mining and information fu-sion from social media and of the new applications and frameworks that are currently appearing under the "umbrella" of the social networks, social media and big data paradigms.

Keywords : Machine Learning, Web, Social Media, Networks, Big Data

I. INTRODUCTION

Big Data, which is characterized by large volume, variety, velocity, openness, inappropriate structure, and visualization among others. Big Data is set to play a major role in various domains such as science, research, engineering, medicine, healthcare, finance, business, and ultimately society itself [2]. It can be used for analyzing and forecasting business trends, profit, and loss, and identifying real-time road traffic conditions, healthcare, weather information, and so on.

Big Data is generally characterized by the 3Vs: variety, volume, and velocity. Variety refers to the nature and structure of the information that constitute the Big Data. Velocity refers to the frequency of data generation as well as the dynamic aspects of the data. Variety refers to the multimodal nature of data such as different data schemas of data sources, structured data like anthologiesa, and unstructured data like sensor signals .

Data volume and the multitude of sources have experienced exponential growth, creating new technical and application chal-lenges; data generation has been estimated at 2.5 Exabytes (1 Ex-abyte = 1.000.000 Terabytes) of data per day . These data come from everywhere: sensors used to gather climate, traffic and flight information, posts to social media sites (Twitter and Facebook are popular examples), digital pictures and videos (YouTube users upload 72 hours of new video content per minute, transaction records, and cell phone GPS signals, to name a few. The classic methods, al-growths, frameworks, and tools for data management have become both inadequate for processing this amount of data and unable to of-fer effective solutions for managing the data growth. The problem of managing and extracting useful knowledge fromthese data sources is currently one of the most popular topics in computing research . In this context, **big data** is a popular phenomenon that aims to provide an alternative to traditional solutions based on databases and data analysis. Big data is not just about storage or access to data; its solutions aim to analyse data in order to make sense of them and exploit their value. Big data refers to datasets that are terabytes to petabytes (and even exabytes) in size, and the massive sizes of these datasets extend beyond the ability of average database software tools to capture, store, manage, and analyse them effectively. The concept of big data has been defined through the **3V model**, which was defined in 2001 by Laney as: "high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making". More recently, in 2012, Gartner updated the def-inition as follows: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization". Both definitions refer to the three basic features of big data: **V**olume, **V**ariety, and **V**elocity. Other organizations, and big data practitioners (e.g., researchers, engineers, and so on), have extended this 3V model to a 4V model by including a new "V": **V**alue. This model can be even extended to 5Vs if the concepts of **V**eracity is incorporated into the big data definition. Summarizing, this set of ∗V-models provides a straightforward and widely accepted definition related to what is (and what is not) a big-data-based problem, application, software, or framework. These concepts can be briefly described as follows -**V**olume: refers to large amounts of any kind of data from any different sources, including mobile digital data creation devices and digital devices. The benefit from gathering, processing, and analyzing these large amounts of data generates a number

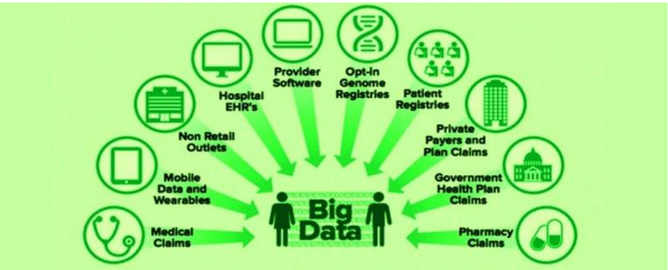

Figure 1. Big Data Real-Time Processing

Further, the processing, availability or acquisition of Big Data can be classified into different categories, including batch processing, real-time processing, and hybrid processing. Batch processing is an efficient.

II. BIG DATA DEFINITION AND CHARACTERISTICS

2.1 Big Data Definition and Growth

The term “Big Data” has several definitions, as it is context specific. In the context of cellular networks, it can be defined as the massive amount of assorted data that can be obtained from the mobile network [21]. From the modern industry point of view, there is another definition that considers big data as the high-volume, high velocity, and high-variety information that demands cost effective decision-making and advanced information processing for enhanced insight [22]. The most general way to define big data is when we weigh its properties to the pre-existing technology as the word “big” is a relative word that exists only for the time being, but when we take into consideration the fast advances in technology, then we can say it is the amount of data that is beyond today’s technology capabilities to store, manage, and process in an efficient and easy way [23]. Since there is no definition for big data that can be labeled as “the universal definition,” it might be better to illustrate its features, and discuss each one to reach a better understanding [10].

Big Data Characteristics

In 2001, Gartner’s analyst Doug Laney dubbed the 3V model that describes big data [24]. The proposed model aimed to illustrate the challenges and show the opportunities introduced by the increase in data. This model was used for the following 10 years by IBM, some Microsoft research departments, and many enterprises [6]. Laney’s model was updated later by many researchers. The new models included additions such as value, veracity, or both, thus adding a fourth or a fifth dimension [10, 22]. Below is an explanation for the Vs indicated earlier.

Table 1: Different Various of big data

BIG DATA - CHARACTERISTICS								
No.of Vs	Volume	Velocity	Variety	Veracity	Value	Variability	Volatility	Validity
Vs3	✓	✓	✓					
Vs4	✓	✓	✓	✓				
Vs5	✓	✓	✓	✓	✓			
Vs6	✓	✓	✓	✓	✓	✓		
Vs7	✓	✓	✓	✓	✓		✓	✓

BIG DATA Vs = F1……….Fn
= Vs3……..Vs7

Vs3	Volume + Velocity + Variety
Vs4	Volume + Velocity + Variety + Veracity
Vs5	Volume + Velocity + Variety + Veracity +Value
Vs6	Volume + Velocity + Variety + Veracity +Value + Variability
Vs7	Volume + Velocity + Variety + Veracity +Value + Volatility + Validity

1) Volume

It might be the simplest aspect for the reader to grasp, as the name implies, it is a representation of the data size. It is considered the most significant and distinctive feature of big data with projections for the monthly mobile data that exceed the 16 exabytes threshold by 2018 (it is worth mentioning that one exabytes can hold 3000 times the contents of the Congress library). Such an immense size of data has forced scientists to reconsider the way they think about size management through storage and processing with the necessity to facilitate the way the data are analysed [10, 25, 26].

2) Variety

There are a variety of sources that can generate data in a network, hence we would have a range of data types that need to be processed, stored, and analysed. Big data can incorporate many data types. These data types can be structured, such as e-mail messages, Twitter tweets, or Facebook contents, etc.; semi-structured, such as log files data from a webpage; and unstructured, such as call centre notes, customer feedback, audio, graphics, natural language data, and hybrid data [22, 27]. The authors of [28] regarded variety as the biggest obstacle facing the effective use of large data volumes.

3) Velocity

Velocity is an indication of the speed of the data when being generated, streamed, and aggregated [28]. It can also refer to the speed at which the data are analysed, whether it is batch processed (for historical data) or stream processed (for real-time generated data) [10]. Currently, data are being generated at rapid speeds, every minute YouTube has 72 hours of video uploads [6]. Depending on the research area, the researchers are required or expected to adopt a specific model incorporating certain characteristics that describe big data. The basic 3V model answers some obvious questions like, how fast is the data? How big is it? And how diverse it is? However, depending on 3 the problem space, the researchers might have other question(s) to answer, hence another term or V can be added. For example, is this data of any value? How long can we consider this an accurate and valid data? Since we are conducting a survey, we find it compelling to briefly introduce other Vs as well. Throughout our survey, we have discovered up to 7V models, as shown in Table **1**.

4) Value

Value is used to measure the data’s usefulness when it comes to decision making [28], it is also defined as how much added-value is brought by the collected data to the intended process, activity, or predictive analysis/hypothesis [26]. Based on the recognition of the great economic and social value within the data, the Obama administration announced a $200 million big data initiative [9, 29].

5) Veracity

This refers to the authenticity and trustworthiness of the collected data against unauthorized access and manipulation [26, 30]. For example, big data might include Facebook posts,tweets, or other social media activities. We cannot label, however, everything in the social media networks as trustworthy [31]. Data related to security applications might include certain encrypted files or quarantined data, similar files are unworthy to process, as it would consume time and resources with no useful output.

6) Volatility

Recalling the retention policy where structured data is destroyed once the retention period expires, big data is not an exception of this policy, this is due to the fact that old data would become a financial burden over time in terms of security and storage [31]. Thus, volatility is used as an indication for the period in which the data can still be regarded as valid and for how long that data should be kept and stored, and this would identify at what point the data becomes irrelevant to the current analysis [32].

7) Validity

This might appear similar to veracity; however, the difference is that validity deals with data accuracy and correctness regarding the intended usage. Thus, certain data might be valid for an application but invalid for another, and this is why a certain amount of relationship verification between data is required to some extent [31]. This can be of major significance. For example, in medical applications where designing a specific healthcare network (e.g. Body Area Network) may take into consideration specific parameters (e.g. allergy to certain materials and symptoms, etc.).

8) Variability

This refers to the inconsistency of the data. This is due to the absence of centralized control over the high number of distributed autonomous data sources [48]. Other researchers refer to the variability as the consistency of the data over time [30].

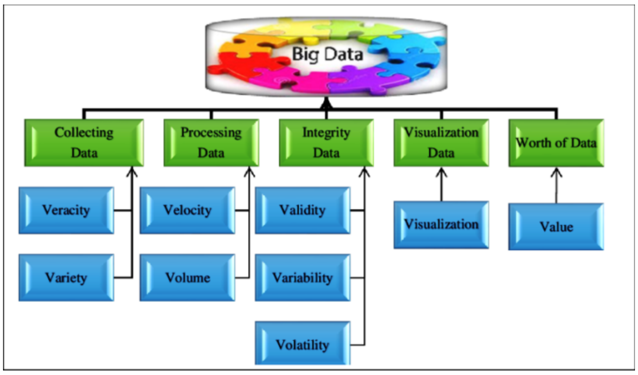

Figure 2. Big Data Processing

2.2 BIG DATA ANALYTICS

The most crucial part of the process of dealing with big data is not the generation or the collection; in fact, it is the processing, as this is how we can benefit from the collected data, by means of getting a decision or predicting a value. The process of examining the hidden patterns and ambiguous correlations within big data for the sake of reaching better decisions is called big data analytics [40]. In a quest to acquire meaningful information from big data, one might mistakenly assume that collecting more data means knowing more useful information, however, the accurate way to reflect on this issue is by keeping in mind the fact that the amount of useful information is not proportional to the amount of acquired data [50]. Big data is gathered from a diverse range of external (e.g., social media [51]) and internal sources (e.g., sensor data [51]).

External sources have to be verified to ensure their worthiness and to make sure they are not dirty. Dirty data refers to duplicated, incomplete (like misspelled words), and incorrect data (like wrong readings from a faulty sensor). It should be noted that a decision has to be made on whether this data should be cleaned or not [52].

Figure 3. Big Data Runtime Processing

Seven Vs Another concept that needs to be addressed to complete this research is the discussions that were taking place between researchers on the seven Vs of big data. There are many scholars debating seven Vs, and one of the best examples is Khan, Uddin, and Gupta's paper [10]. This paper concisely revealed the seven Vs. The seven characteristics according to them are: Volume, Velocity, Variety, Veracity, Value, Validity, and Volatility. The two characteristics added are Validity and Volatility where:

examples is Khan, Uddin, and Gupta's paper [10]. This paper concisely revealed the seven Vs. The seven characteristics according to them are: Volume,

Velocity, Variety, Veracity, Value, Validity, and Volatility. The two characteristics added are Validity and Volatility where:

Validity is the accuracy and correctness of the data with respect to the intended use. Although it sounds so similar to veracity, they are two different concepts. A data set may have no problem with veracity, but it may still not be valid. In other words, without validating it, we cannot simply take a data set, and trust it.

Volatility is the responsibility for rapid and unexpected transitions. There are many businesses that have openly admitted that they are not storing older data that have no value. Online companies, for example, may not want to keep older consumer purchasing history, since the warranty may expire. It is important to ensure the volatility of a data set to allow complete reliability of the final outcome.

3. PROPOSED METHODOLOGY FOR CHARACTER BY USING BIG DATA ANALYTICS

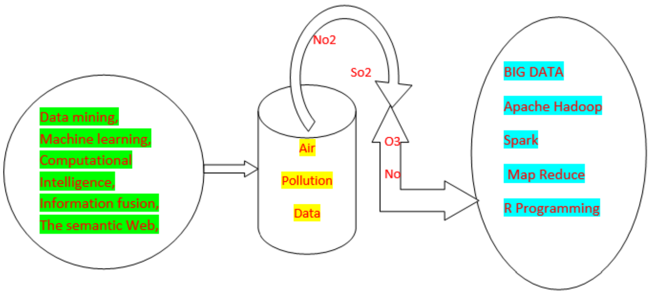

Figure 3. Big Data Model

The following figure (3) the Proposed Methodology for character by using Big data Analytics. Proposed Methodology for Big data

III. RESULT AND DISCUSSION

once addressing volume, velocity, variety, variability, veracity, and visualization which takes a lot of time, effort, and resources you want to be sure your association is getting worth from the data. The big data is data that contains greater variety, arriving in increasing volumes and with more velocity. This is also known as the three Vs. Put simply, big data is larger, more complex data sets, especially from new data sources.Five Primary Characteristics of High Quality Data Accuracy, Completeness, Validity, Consistency, Timeliness. There is one "V" that we stress the importance of over all the others veracity. Data veracity is the one area that still has the potential for improvement and poses the biggest challenge when it comes to big data.

IV. CONCLUSION

The proposed methodology has analyzed from the past data and air pollution prediction using big data environment which has stored in the National Climatic Data Centre (NCDC) is an efficient solution. Map reduce is a framework for highly parallel and distributed systems across large dataset. This type of technology used to analyze large datasets has the potential for significant enhancement to the air pollution.

V. REFERENCES

[1]. Nagham Saeed*, Laden Husamaldin "Big Data Characteristics (V's) In Industry" Iraqi Journal Of Industrial Research (IJOIR),Vol 8,No.1(2021) ,Journal Homepage: Http://Ijoir.Gov.Iq

[2]. Talat ZAREE, Ali Reza HONARVAR,"Improvement Of Air Pollution Prediction In A Smart City And Its Correlation With Weather Conditions Using Metrological Big Data" Turk J Elec Eng & Comp Sci (2018)26:1302{1313-TUBITAK_Doi:10.3906/Elk-1707-99 Turkish Journal Of Electrical Engineering & Computer Sciences. Http://Journals.Tubitak.Gov.Tr/Elektrik

[3]. Yue Shan Chang, Kuan-Ming Lin, Yi-Ting Tsai. "Big Data Platform For Air Quality Analysis

And Prediction” The 27th Wireless And Optical Communications Conference (WOCC2018) 978-1-5386-4959-6/18/$31.00 ©2018 IEEE,Https://Spark.Apache.Org.Https://Www.Lora-Alliance.Org.,Https://Spark.Apache.Org,Https://Www.Cwb.Gov.Tw/V7/, Https://Www.Tensorflow.Org

[4]. S.Suganya, Dr.T.Meyyappan “Adaptive Deep Learning Model For Air Pollution Analysis Using Meteorological Big Data” Date Of Conference: 16-17 December 2021Date Added To IEEE Xplore: 28,January 2022 ISBN Information: INSPEC Accession Number-21702330 , DOI:10.1109/C2I454156.2021.9689298,Publisher: IEEE,Conference Location: Bangalore,India

[5]. Z.Ghaemi, A. Alimohammadi, And M. Farnaghi. "Lasvm-Based Big Data Learning System For Dynamic Prediction Of Air Pollution In Tehran."Environmental Monitoring And Assessment 190, No. 5 (2018): 1-17.

[6]. Wojciech Zaremba, Ilya Sutskever, And Oriol Vinyals. "Recurrent Neural Network Regularization." Arxiv Preprint Arxiv:1409.2329 (2014).

[7]. Cheng Fan, Jiayuan Wang, Wenjie Gang, And Shenghan Li."Assessment Of Deep Recurrent Neural Network-Based Strategies For Shortterm Building Energy Predictions." Applied Energy 236 (2019): 700-710.

[8]. Laith Abualigah, Ali Diabat, Seyedalimirjalili, Mohamed Abd Elaziz, And Amir H. Gandomi. "The Arithmetic Optimization Algorithm."Computer Methods In Applied Mechanics And Engineering 376 (2021): 113609.

[9]. Samir Khatir, Samir Tiachacht, Cuong Le Thanh, Emad Ghandourah, Seyedalimirjalili, And Magd Abdel Wahab. "An Improved Artificial Neural Network Using Arithmetic Optimization Algorithm For Damage Assessment In FGM Composite Plates." Composite Structures 273(2021): 114287.

[10]. Copernicus Climate Change Service Climate Data Store (CDS). Fifth Generation Of ECMWF Atmospheric Reanalysis Of The Global Climate Https://Ads.Atmosphere.Copernicus.Eu/Cdsapp#!/Home

Cite this article as :

S. Suganya, T. Meyyappan, "Big Data and Seven (7) V's Characteristics in Industry", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 9 Issue 6, pp. 630-636, November-December 2022.

Journal URL : https://ijsrst.com/IJSRST2296109