GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

7 DIMENSIONS OF BIG DATA ANALYTICS

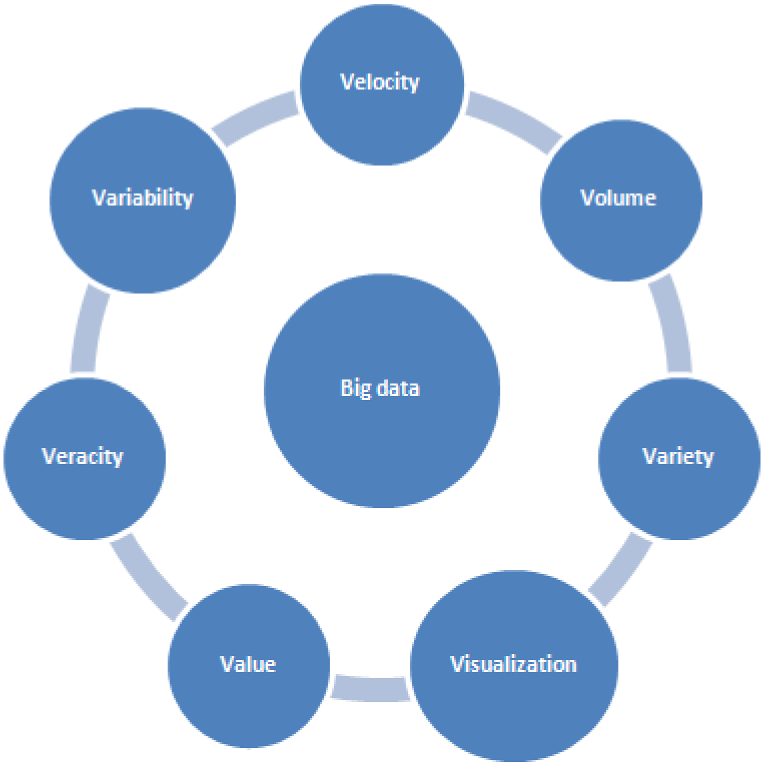
Prof. Sachchidanand Nimankar*[1] & Prof. Sushant Dagare[2]
[*1]Assistant Professor, Mechanical Department, SSPM's College of Engineering, Kankavli, India
[2]Assistant Professor, Computer Science Department, SSPM's College of Engineering, Kankavli, India

ABSTRACT

Big data is a term for massive data sets having large, more diversified and very complex structure along with the difficulties like storage, analyze and visualize for further processes or results. The process of research into massive amounts of data to identified hidden patterns and secret correlations named as big data analytics. These useful information for companies or organizations with the help of gaining richer and deeper insights and getting an advantage over the competition. For this reason, big data implementations need to be analyzed and executed as accurately as possible. This paper attempts to offer a broader definition of big data that captures its other unique and defining characteristics and represents an overview of 7 dimensions of big data.

Keywords: *Volume, Velocity, Variety, Value, Visualization, Variability, Veracity*

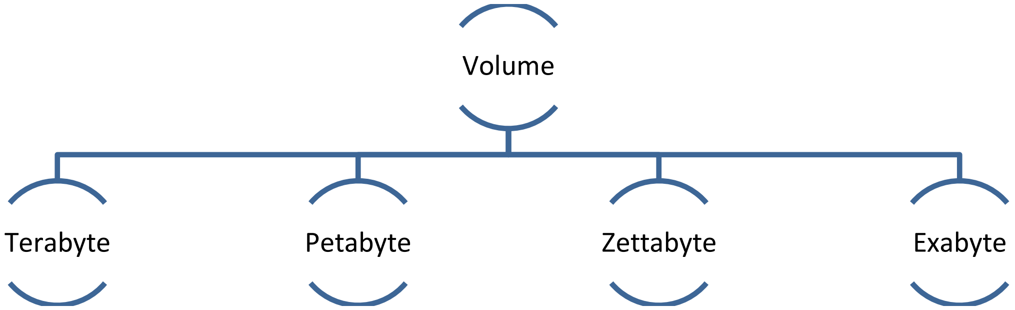
I. INTRODUCTION



Big data is not just about storage or access to data but its solutions analyze data in order to make a sense of the mand increase their value. Big Data are datasets whose size is beyond the ability of conventional database software toolsto capture, store, manage, and analyze.[1]Applications requiring effective analyses of large datasets are widely recognized today's world. Such applications include health care analytics (e.g., personalized genomics), business process optimizations and social-network-based recommendations.[2]Size is the first, and at times, the only dimension that leaps out at the mention of big data.. The rapid evolution and adoption of big data by industry has jumped the discourse to popular outlets, forcing the academic press to catch up. Academic journals in numerous disciplines, which will benefit from a relevant discussion of big data, have yet to cover the topic. The paper's primary focus is on the 7 dimensions of big data. This paper also reinforces the need to devise new tools for predictive analytics for structured big data. The statistical methods in practice were devised to infer from sample data. The heterogeneity, noise, and the massive size of structured bigdata calls for developing computationally efficient algorithms that may avoid big data pitfalls, such asspurious correlation.

II. VOLUME

Volume is simply defined as the large data-sets consisting of terabytes, petabytes, zetta bytes of data – or even more large scale and the sheer volume of data is a big challenge in its own right. Facebook daily generates over 500 terabytes of data, and Wal-Mart collects more than 2.5 petabytes of data every hour from its customer transactions.[3] Datasets over one terabyte is considered to be big data. One terabyte stores as much data as would fit on 1500 CDs or 220 DVDs and it is enough to store around 16 million Facebook photographs. Report says that Facebook processes up to one million photographs per second. One petabyte equals1024 terabytes. Earlier estimates suggest that Facebook stored 260billion photos using storage space of over 20 petabytes. Definitions of big data volumes are relative and vary by factors, such as time and the type of data [4].In the coming future, as the data sizes continue to grow and the domains of these applications diverge, these systems will need to adapt to leverage application-specific optimizations.[2]The amount of data on the web is measured in Exabyte (10^{18}) and zettabytes (10^{21}). By 2025, the forecast is that the Internet will exceed the brain capacity of everyone living in the whole world.[1]

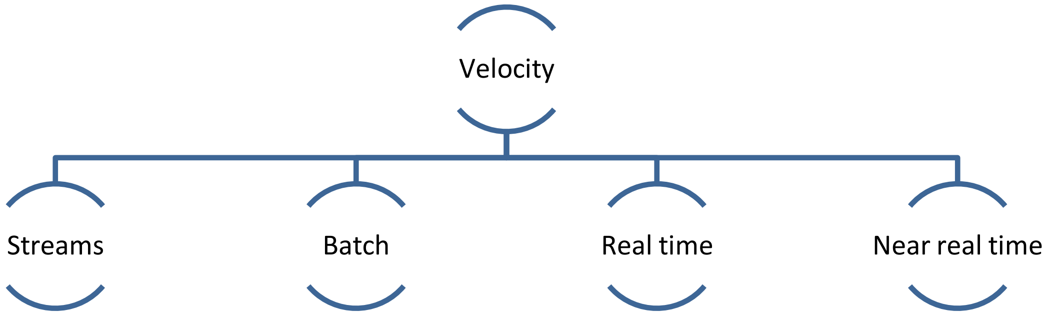


III. VELOCITY

The Velocity is the high rate of data inflow with non-homogenous structure. For instance, Wal-Mart processes more than a million transactions each hour [3].The big challenge of velocity comes with the importance to manage the high influx rate of homogenous/non-homogenous structured/unstructured data, which results in either creating new data or updating the existing data.The explosion of digital devices such as smart phones and sensors has led to an unprecedented rate of data creation and is driving a growing need for real-time analytics and evidence-based planning.[4]The data's contents are constantly changing through the absorption of complementary data collections, the introduction of previous data or legacy collections, and the different forms of streamed data from

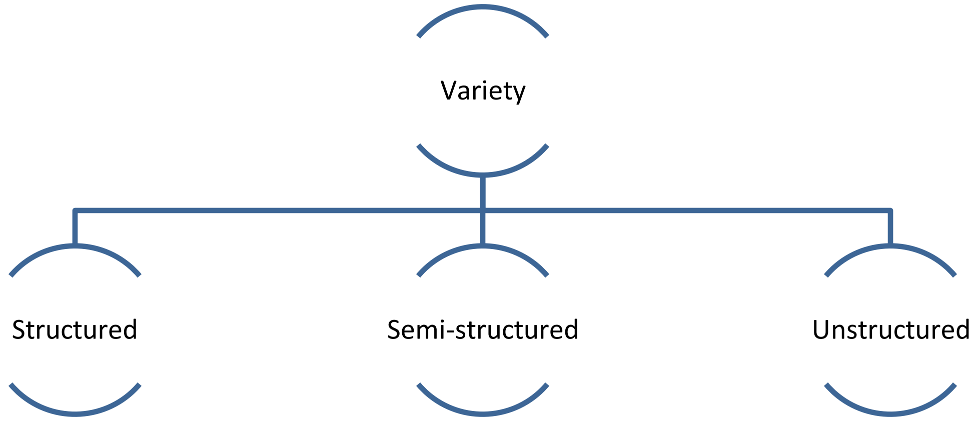
multiple sources.[5]Indeed it is not just the velocity of the incoming data that is the issue it is possible to stream fast-moving data into bulk storage for later batch processing, for example. The importance lies in the speed of the feedback loop, taking data from input through to decision [1]Recent conservative studies estimate that enterprise server systems in the world have processed 9.57×10^{21} bytes of data in 2008. This number is expected to have doubled every two years from that point. As an example, Wal-Mart servers handle more than one million customer transactions every hour, and this information is inserted into databases that store more than 2.5 petabytes of data the equivalent of 167 times the number of books in the Library of Congress. The Large Hadron Collider at CERN will produce roughly 15 peta bytes of data annually enough to fill more than 1.7 million dual-layer DVDs per year. Each day, Facebook operates on nearly 500 terabytes of user log data and several hundreds of terabytes of image data. Every minute, 100 hours of video are uploaded on to YouTube and upwards of 135,000 hours are watched. Over

28,000 multi-media (MMS) messages are sent every second. Roughly 46 million mobile apps were downloaded in 2012, each app collecting more data. Twitter serves more than 550 million active users, who produce 9100 tweets every second. eBay systems process more than 100 petabytes of data every day. In other domains, Boeing jet engines can produce10 terabytes of operational information for every 30 min of operation. This corresponds to a few hundred terabytes of data for a single Atlantic crossing, which, if multiplied by the 25,000 flights each day, highlights the data footprint of sensor and machine-produced information.[2]



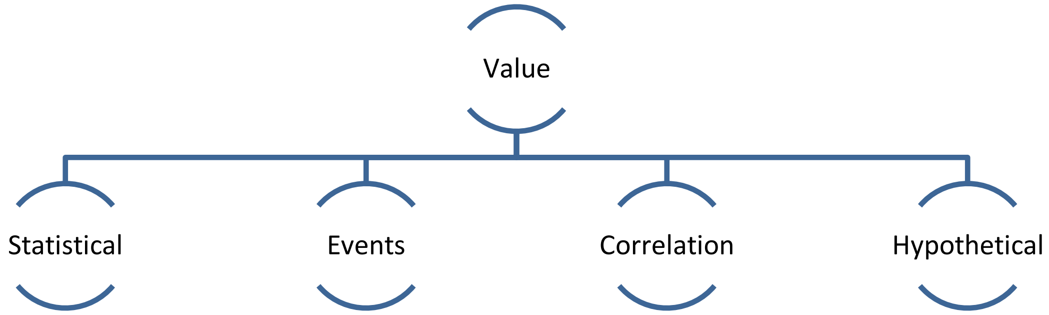
IV. VARIETY

Variety is simply defined as the multiple data formats with structured and unstructured text, image, multimedia contents like audio, video, sensor data like noise.[3]Diverse and dissimilar nature of data is a Big challenge. The enormous volume of data generated is not consistent nor does it follow a specific template or format rather it is captured in diverse forms and from diverse sources like messages (text, email, tweets, blogs etc), user generated contents, transactional data (web logs, business transactions etc), scientific data (data coming from data-intensive experiments – genome and healthcare data etc), web data (images posted on social media, sensor data readings etc) and much more. These different forms and quality of data clearly indicate that heterogeneity is a natural property of Big Data and it is a big challenge to understand and manage such data.Technological advances allow firms to use various types of structured, semi-structured, and unstructured data. Structured data, which constitutes only 5% of all existing data.[4]



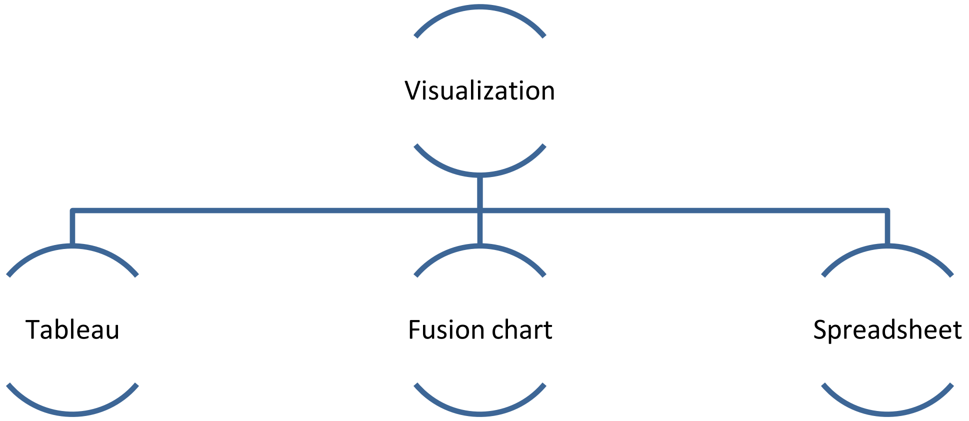
V. VALUE

Value is the extracting knowledge from vast amounts of structured and unstructured data without loss, for end users.[3]Big data researchers consider value as an essential feature, as somewhere within that data, there exist valuable information. Extracting such golden data or high-valued data, though most of the pieces of data independently may seem insignificant. Oracle introduced Value as a defining attribute of big data. Based on Oracle’s definition, big data are often characterized by relatively “low value density”. That is, the data received in the original form usually has a low value relative to its volume. However, a high value can be obtained by analyzing large volumes of such data.[4]Value is the most important characteristic of any Bigdata based application, because it allows to generate useful business information.[5]This value falls into two categories viz. analytical use (replacing/supporting human decision, discovering needs, segmenting populations to customize actions) and enabling new business models, products and services[1]



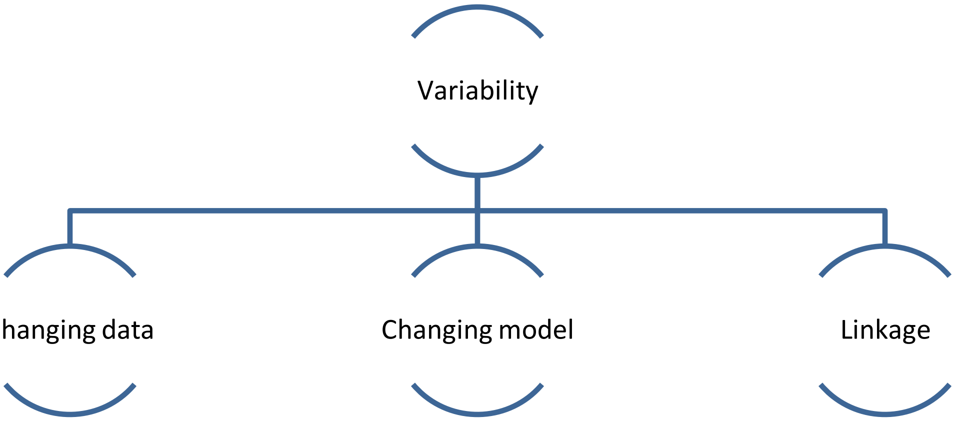
VI. VISUALIZATION

Visualization is nothing but presenting the data in a manner that is readable. Visualizing the data is about representing key information and knowledge more instinctively and effectively through using different visual formats such as in a pictorial, tabular or graphical layouts.[3]Tableau, spreadsheet are popularly used for data visualization and presentation. Which are capable of transforming large and complex datasets into spontaneous descriptions. Based on these interactive results, one can visualize search relevance and quality to monitor the latest customer feedback and conduct sentiment analysis.



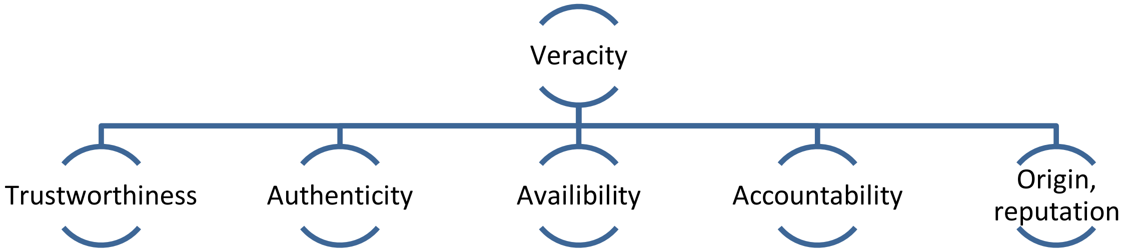
VII. VARIABILITY

Variability is the data whose meaning is constantly changing. Among the seven pillars of Big Data, variability is another extremely essential feature but is often confused with variety.[3] Variability is also related in performing sentiment analyzes. For instance, suppose one organization, repository stores and generates many different types of data. At the same time, if from these different types of data, one of them is brought to use for mining and making sense out of it but every time the data offers a different meaning, this is the variability of data, whose meaning is constantly and rapidly changing.SAS introduced this term. Variability refers to the variation in the data flow rates. Often, big data velocity is not consistent and has periodic peaks and troughs.[4]



VIII. VERACITY

Veracity is the increasingly complex data structure, anonymities, imprecision or inconsistency in large data-sets. This is not merely about data quality – it is more about understanding the data, as there are integral discrepancies in almost all the data collected. Veracity feature measures the accuracy of data and its potential use for analysis[3].IBM coined Veracity as the unreliability inherent in some sources of data. For example, customer sentiments in social media are uncertain in nature, since they entail human judgment. Yet they contain valuable information. [4]Veracity refers to the correctness and accuracy of information. Behind any information management practice lie the core doctrines of data quality, data governance, and meta data management, along with considerations of privacy and legal concerns.[5]Veracity is what is conform with truth or fact, or in short, Accuracy, Certainty, Precision. Uncertainty can be caused by inconsistencies, model approximations, ambiguities, deception, fraud, duplication, incompleteness, spam and latency. Due to veracity, results derived from Big data cannot be proven; but they can be assigned a probability.[1]



IX. CONCLUSION

The objective of this paper is to describe, review the 7 dimensions of big data. The paper first defined what is meant by big data to consolidate the divergent discourse on big data. We presented various definitions of big data, highlighting the fact that dimensions, such as velocity and variety, volume, visualization, veracity, volume and variability are equally important. However, big data technologies enabled businesses to adopt sentiment analysis to gather useful insights from millions of opinions shared on social media. Since big data are noisy, highly interrelated, and unreliable, it will likely lead to the development of statistical techniques more readily apt for mining big data while remaining sensitive to the unique characteristics. Going beyond samples, additional valuable insights could be obtained from the massive volumes of less ‘trustworthy’ data

.

REFERENCES

1. *CheikhKacfahEmani, Nadine Cullot, Christophe Nicolle, "Understandable Big Data: A survey," in © 2015 Elsevier Inc.,computersciencereview17(2015)70‑81.*
2. *KarthikKambatlaa,∗, GiorgosKollias b, VipinKumarc, AnanthGramaa, "Trends in big data analytics," in J. Parallel Distrib. Comput. 74 (2014) 2561–2573*
3. *UthayasankarSivarajah , Muhammad Mustafa Kamal, ZahirIrani, VishanthWeerakkody, "Critical analysis of Big Data challenges and analytical methods," in Journal of Business Research 70 (2017) 263–286*
4. *Amir Gandomi∗, MurtazaHaider, "Beyond the hype: Big data concepts, methods, and analytics," in International Journal of Information Management 35 (2015) 137–144.*
5. *GemaBello-Orgaza,JasonJ.Jungb,∗,DavidCamachoa,"Socialbigdata:Recentachievementsandnewchallenges," in science direct Information Fusion 28 (2016) 45–59.*