5V's of Big Data Attributes and their Relevance and Importance across Domains

Vinaya Keskar, Jyoti Yadav, Ajay H. Kumar

Abstract - "Data is an ocean of Universal Facts". Big data once an emergent technology of study is in its prime with immense potential for future technological advancements. A formal study in the attributes of data is essential to build robust systems of the future. Data scientists need a basic foot hold when studying data systems and their applications in various domains. This paper intends to be THE go-to resource for every student and professional desirous to make an entry in the field of Big Data. This paper has two focus areas. The first area of focus is the detailing of the 5 V attributes of data i.e. Volume, Variety, Velocity, Veracity and Value. Secondly, we will endeavor to present a domain wise independent as well as comparative of the correlation between the 5 V's of Big Data. We have researched and collected information from various market watch dogs and concluded by carrying out comparatives which are highlighted in this publication. The domains we will mention are Wholesale Trade Domain, Retail Domain, Utilities Domain, Education Domain, Transportation Domain, Banking and Securities Domain, Communication and Media Domain, Manufacturing Domain, Government Domain, Healthcare Domain, etc. This is invaluable information for Big Data system designers as well as future researchers.

Keywords*: *5V's, Big Data, Data Attribute, Quantity, Reliability, Volume, Variety, Velocity, Veracity, Value, Worth

I. INTRODUCTION

"Data is an Ocean of Universal Facts", statement defines that all facts of the universe can be said to be data. Decisions that need to be taken based on the facts must ensure that the inferential value of the relevant facts are appropriately computed or ascertained. This can be done by carrying out data analytics. It is only after the data is processed, one can take at intelligent and wise decisions. The rift between the increasing quantity of data and the derived information there off is getting wider, deeper and more complex with every passing minute. Now the industry experts do not even call it data anymore and are forced to call it "BIG DATA".

The objectives of business houses are always to take decisions which will lead to profitable growth in business and improved customer satisfaction. The transaction cycle will always start at data that leads to data analytics for wise decisions and it will always lead to taking decisions based on analyzing patterns and behaviors of both humans as well as systems. A journey from collecting data, setting objectives and then taking wise decisions through building knowledge quotient is a sequence of steps that leads to data analytics for wise decision, as in figure-1.

Revised Manuscript Received on September 05, 2020.

Vinaya Keskar, Research Student, Department of Computer Science, Savitribai Phule Pune University (SPPU), Pune, Maharashtra, India.

Dr. Jyoti Y. Yadav, Assistant Professor, Department of Computer Science, Savitribai Phule Pune University, Pune, Maharashtra, India.

Dr. Ajay H. Kumar, Research Guide, Department of Computer Science, Savitribai Phule Pune University, Pune, Maharashtra, India.

Before we deep dive into the taxonomy of the attributes of big data it is necessary to get a brief idea about the flow of data to information leading to informed decision making.

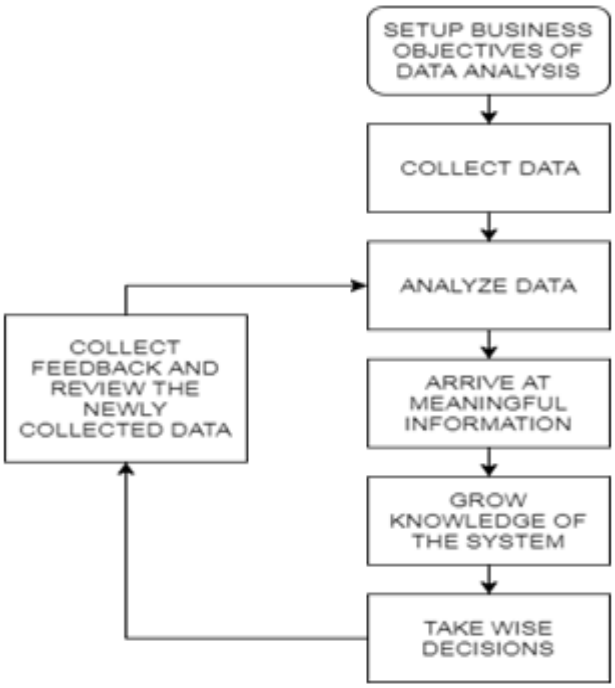


Figure-1: The sequence of steps involved in data analysis which lead to taking 'wise' decisions.

The start has to be very well defined and articulated. To elaborate Big Data Analytics definition - This is a pipeline of acquisition and recording; extraction cleaning and annotation; integration, aggregation and representation; analysis and modeling; interpretation. This leads us to conclude that a study in attributes of data is critical and the most important aspect of the complete life cycle. It is the foundation of all things which can be classified as knowledge. Also, no amount of analysis can by-pass or justify any loss / lack of understanding of the data leading to decisions. It is very important that knowing the attributes of data and these are key imperatives to this world of Big Data. There are various proposed attributes of Big Data based on various thought processes. Different companies have deployed various ideations for defining the key attributes of data. This survey is centered around exploring these attributes with respect to their definition in different Business domains, their importance and inferential value. In addition, it also compares the importance and challenges of the attributes across the top business domains.

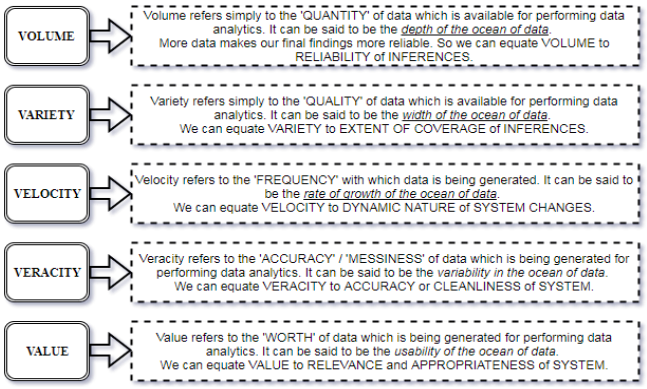
II. 5V'S OF BIG DATA

In 2001 Meta Group publication and Gartner analyst Doug Laney had introduced the 3 V's of Data Management, defining the 3 main components of data as Volume, Velocity and Variety [1]. This can be considered as the first and most important step towards the revolution called Big Data attributes for Analytics as it paved a way in a direction to arrive at more ergonomic and reliable outcomes. The outcomes would now be more accurate and lead to better profitability given that the inferences arrived at, were in close proximity to the true picture or reality.The growth of the ocean of data soon outlived these three attributes and there was a need to add more attributes for a more robust analysis to happen.

IBM is attributed to the addition of the 4th V to the attribute chain. It introduced Veracity and defined it as: "Uncertainty of the data"- The quality or trustworthiness of the data. One of the tools that helps to handle big data's veracity is to discard "noise" and transform the data into trustworthy insights." [2]

As the world moves ahead and the ocean of data continues to grow beyond boundaries, the hunger of Man to grow in leaps and bounds has now led to the need for two new V's namely Variability and Visualization. However, it is not the endeavor to explore these attributes.

Kapil, Agarwal and Khan (2016) in an independent study mentions about the characteristics of big data. The study presented that Big data is a new idea, and it has got numerous definitions from researchers, organizations, and individuals. In 2001, industry analyst Doung Laney (currently with Gartner), articulated the mainstream of definition of big data regarding in terms of three V's; Volume, Velocity, and Variety [13]. SAS (Statistical Analysis System) has added two additional dimensions i.e. Variability and complexity. Further, Oracle has defined big data in terms of four V's i.e. Volume, Velocity, Variety and Value. Furthermore, Oguntimilehin A, presented big data in terms of five V's Volume, Velocity, Variety, Variability, Value and a Complexity [14]. In 2014, Data Science Central, Kirk Born has defined big data in 10 V's i.e. Volume, Variety, Velocity, Veracity, Validity, Value, Variability, Venue, Vocabulary,
Vagueness [13]



Now, this 5 V's of data can be summarized in figure-2.

Figure-2: The characteristics of Big Data Attribute (5-V)

VOLUME: Volume is the first data attribute and is the foundation of all other attributes. It refers to the measure of the quantity of data available. The more the data that is available the more clarity will exist in the other attributes of the data. In case the data available is "very small" [very small in the world of big data would still be in Tera Bytes at times] it would not be possible to express other attributes of the data with a high degree of belief. We would be forced to have more assumptions to substantiate our facts or hypothesis. In short:

VOLUME ☞ QUANITY ☞DEPTH ☞ RELIABILITY OF INFERENCES

VARIETY: This is the proverbial spread of the data types that can be analyzed. It also refers to the breadth or the extent of coverage. It can be said to be the flavor of data which exists either in terms of structured, semi-structured or a completely unstructured set as the input on which further analytics are to be performed. The quality or inconsistencies defines the value that can be generated from using such data sets. In short:

VARIETY ☞ SPREAD ☞ WIDTH ☞ SCOPE OR EXTENT OF COVERAGE

VELOCITY: This attribute defines the rate of growth of data. In this age of internet and technological advances where not only humans but systems are smart enough to generate data and define personal preferences to gather data without human intervention, the rate of growth of the 'pool' or 'ocean' of data both in terms of the depth as well as width is astronomical.

VELOCITY ☞ SPREAD AND WIDTH INCREASE ☞ RATE OF GROWTH OF DATA

VERACITY: With the growth of the volumes of data and the high velocities with which the data is being assimilated, there is bound to be a lot of errors or duplication or corruption of data. This can also be said to be the noise in the system which is collateral damage to the volume and velocity of big data gamut.

VERACITY ☞ACCURACY ☞MESSINESS ☞ RELIABILITY OF DATA

VALUE: Last but the most important is the Value attribute. This is the icing on the cake. Without this all the other attributes are just meaningless. High volumes of good quality data shall be well analyzed for value and meaningful proposition for direct relevance to the business requirements of wise decisions for business growth, profitability and customer satisfaction.

VALUE ☞ WORTH ☞ RELEVANCE ☞ DIRECT VALUE-ADD TO THE BUSINESS

III. ROADMAP OF 5-V OF BIG DATA ATTRIBUTE

VOLUME ATTRIBUTE:

Volume is defined as "The quantity of generated and stored data. The size of data determines value and potential insight- and whether it can be considered big data or not."[16]

The foundation of Big Data Analytics is dependent on the quantity of data available.

It encompasses everything. To name a few digital data, financial transactions, scientific experiments, genomic data, logs, events, emails, social media, sensors, texts, audio data, medical records, surveillance, images, and videos.

Fremont Rider (1944), Wesleyan University Librarian, publishes The Scholar and the Future of the Research Library. He estimates that American university libraries were doubling in size every sixteen years. Given this growth rate, Rider speculates that the Yale Library in 2040 will have "approximately 200,000,000 volumes, which will occupy over 6,000 miles of shelves…" October 2000 Peter Lyman and Hal R. Varian at UC Berkeley publish "How Much Information?" The study finds that in 1999, the world produced about 1.5 Exabyte of unique information, or about 250 MB's for every man, woman, and child on earth.

June 2008 Cisco releases the "Cisco Visual Networking Index – Forecast and Methodology, 2007–2012" part of an "ongoing initiative to track and forecast the impact of visual networking applications." It predicts that "IP traffic will nearly double every two years through 2012" and that it will reach half a zettabyte in 2012. The forecast held well, as Cisco's latest report (May 30, 2012) estimates IP traffic in 2012 at just over half a zettabyte and notes it "has increased eightfold over the past 5 years."

There are 277,000 Tweets every minute, Google processes over 2 million search queries every minute, 72 hours of new video are uploaded to YouTube every minute, more than 100 million emails are sent every minute, Facebook processes 350 GB of data every minute and 571 new websites are created every minute.

The below chart provides the volumetric distribution of Big Data across various domains.

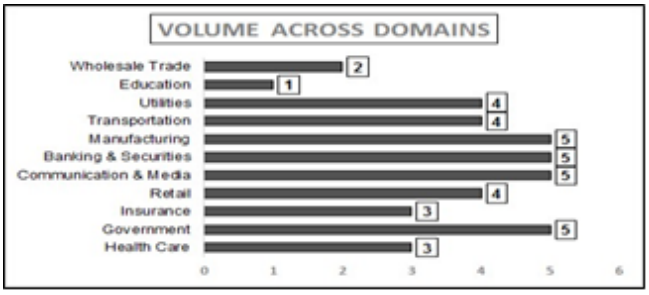


Figure-3 Volume wise distribution of Big Data across domains. [(Ratings: Very high-5, high-4, avg.-3, low-2, very low-1)]

From above study data it is observed that, the challenges with volumes are enormous. This data is available in abundance. This abundance of data creates the below challenges:

- Sampling analysis of data does not allow for a smaller data set as sample size itself is again Big Data. Thus, it cannot be said that sampling will lead to effective and 'easy' analytics for the data.
- Space – The storage and access for such large volume of big data requires storage space, infrastructure, time and speed for analyzing big data. Expecting faster analysis with this big data requires further technological advancements in the field of hardware. Cloud based solutions and shared processing are cost effective solutions to the industry but this is still in the infancy.
- The costs to store and maintain sanctity of such large volumes of data are astronomical. Also, processing of such large data at high speeds requires state of the art systems and tools.
- In addition to this the biggest concern is the privacy and protection of the data. [9]

IV. ARIETY ATTRIBUTE

Variety of data is "The type and nature of the data." This helps people who analyze it to effectively use the resulting insight. The 'data' is generated as by date, time, place, number or *some* textual content. This data is either structured/ semi structured or unstructured. From structured data set, more set of unstructured data sets are generated which results into more opportunities of analysis. For looking into variety of data and quantify the abstract, the industry needs newer algorithms and systems.

Data can now be categorized as:

STRUCTURED DATA: Structured data is the data which resides within a fixed field type. It has a fixed shape, size, format, type, etc. This type of data is to be found in data bases or excel spreadsheets where each record is made up of multiple attributes and the data content at each cell conforms to the record and the attribute type. Examples of structured data are names, phone numbers, ZIP codes, salutations, age, etc.

UNSTRUCTURED DATA: Unstructured data is the next new. This data has non-standard formats or rather where no standard formats exist. The data in its unstructured form cannot be stored in data bases directly or analyzed easily by the conventional tools and methods. The typical examples of unstructured data are photographs, audio clips, video clips, blog entries, forums, social media platforms like Facebook and LinkedIn, presentations, pdf files, web pages, etc. It is estimated that more than 80% data is of the unstructured type.

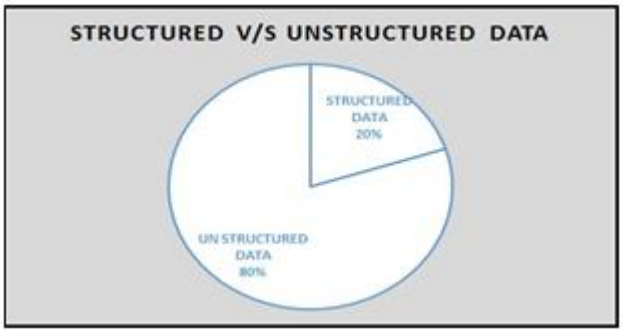


Figure-4: Structured v/s Un Structured Data [7].

In addition, there also exists a mid-class called Semi-Structured data. This semi structured data contains combination of structured and unstructured data. The minute detail study is outside the scope of this paper. Some examples are comma separated value files and unformatted data dumps in databases.

The below chart provides the variety distribution of Big Data across various domains.

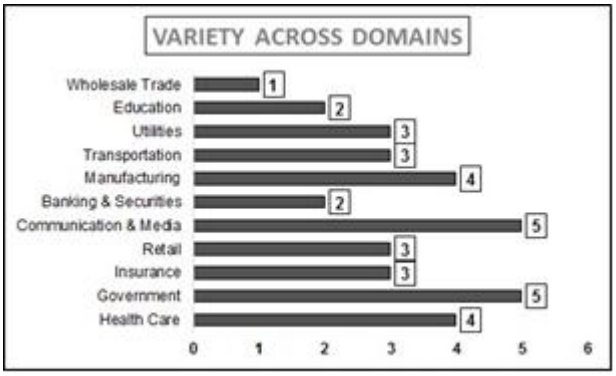


Figure-5: Variety wise distribution of Big Data across domains. [(Ratings: Very high-5, high-4, avg.-3, low-2, very low-1)]

In view of variety wise distribution across domains, the challenges with variety are enormous. This abundance of sub data which is very subjective creates the below challenges:

- A record of data for a single entry can run into hundreds of Tera bytes. So analysis of patterns across multiple objects of a similar category are practically unthinkable or very time consuming and costly.
- The opportunities for error increase many fold as even a few readings which are misplaced will skew the analysis leading to loss of inferential value.
- The vastness and vagueness of data creates a loss of reliability and higher obscurity.
- On the web, 58% of the available documents are XML, among which only one third of XML documents with accompanying XSD/DTD are valid. 14% of the documents lack well-formedness, a simple error of mismatching tags and missing tags that renders the entire XML-technology useless over these documents. [8]

V. ELOCITY ATTRIBUTE

Velocity is defined as "the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development." [16] Velocity not only relates to the speed but also the sources being a vector quantity. For a scalar quantity, a collective quantitative analysis or understanding shall suffice. But velocity being a vector quantity, both the collective quantity and the sources or domains from where the data is being generated needs analysis. This paves way for the opportunities and cost benefit analysis. In domains with lower velocities, initiatives of high value and high importance if only considered, it may not achieve the expected or desired results. When velocity attribute is to be considered then, the velocity at the point of origin of data shall be considered along with the path it takes from the source to the intended destination and the time consumed. Here the bandwidth and technical / technological hardware and software capabilities and availabilities are required so that the data maintains the features of timeliness i.e. being available at the right time for the right purpose. A delay in this will cause a ripple effect and downstream systems will be exponentially impacted.

With the huge volume of generated data, the fast velocity of arriving data, and the large variety of heterogeneous data, the quality of data is far from perfect. [8]

Velocity of the data can be roughly defined in-terms of the below 4 types:

Real Time Data: All data which is created / generated in present time and consumed by downstream or other systems is classified as real time data. For example, a telephone conversation between two or more people, chatting between two people, video conferencing and video calls, live telecast, etc. When a shopper hits the ATM, the bank balance and transactional data have to be processed instantly, or so close that the customer doesn't even notice the delay. This data is very critical as the chances of loss are very high. [10]

Near Real Time Data: Near Real Time data is as the name suggests conforming to real time definition however, there is a delay introduced between the transmitter and the receiver. One can consider the noise in system which leads to delay in transmission or transmission losses. Examples are the time delay in receiving the OTP [one time pin] on one's hand phone for banking and e-commerce based transactions to be completed, time delays in transmission during live broadcasts. Other examples can be where there are multiple live feeds entering the same system and then they are being processed together to provide a single output or specific feed output which forms the input for other downward systems. This is also defined as CEP or Complete Event Processing. This is a part of Operational Intelligence and is a very powerful tool for sales and strategic planning in large organizations where the sensitivity and specificity are of utmost importance in devising smart {*can be compared to S.M.A.R.T. indicating Simple Measurable Achievable, Realistic and Time-bound*} growth plans.

Batch Processing: Batch Processing can be defined as processing of archived or historical data in real time to arrive at patterns. They are not very sensitive and the burden or constraint of time is not of paramount concern. Batch processes run for hours and days on end. Some examples can be activities as simple as payroll processing which run once monthly or batch processes which run as per scheduled tasks or activities on server to much more complex activities like analysis of vast non-parametric data and performing confirmatory data analysis on them using Test of Hypothesis or creating Design of Experiments, etc.

Virtual Time Data: Smart systems and technology has made it possible to invoke communication between interested parties in virtual time. This concept can be expressed as communication between parties when the other party may or may not be available at real time. Thus when the recipient becomes available the message is delivered. Virtual time communication has made it possible for parties to communicate at their time of choosing which is transmitted to the recipients' server spaces waiting for the recipient to retrieve the message. This has led to a blast in the data that is available over the internet. Social media platforms, email, mobile social network software, voice mails, uploading video messages, etc. are all examples in this category.

The below chart provides the velocity distribution of Big Data across various domains.

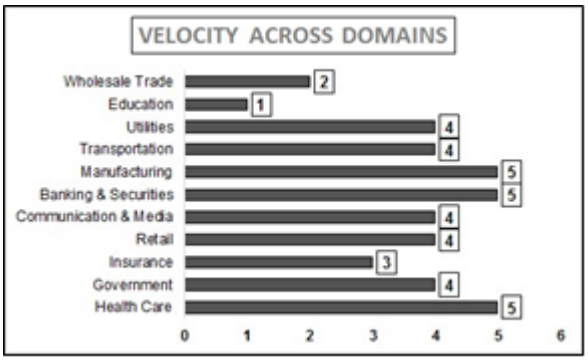


Figure-6: Velocity wise distribution of Big Data across domains. [(Ratings: Very high-5, high-4, avg.-3, low-2, very low-1)]

Given the above study available, the major challenges with velocity are listed below:

- Large volume of generated data and the rate of arrival / transmission of data leads to losses attributable to transmission losses as well as noise which is inherent in the system and is a natural cause. [8]
- Privacy of data and hacking of data which is travelling at such velocities is a very critical and poses a great challenge. To maintain the data integrity redundancy checks or firewall/proxy servers' checks slow down the traffic of data leading to the direct impact on the processing times. Also as the volumes are high if velocity is impacted then the systems in all possibility will choke and crash.
- Infrastructure costs to maintain data velocities and ensuring minimal losses are extremely costly and not easily available. While cloud based infrastructure is a major step forward in this direction, there is a still a long way ahead amidst privacy and protection of data concerns.

VI. ERACITY ATTRIBUTE

Veracity is defined as "The quality of captured data can vary greatly, affecting accurate analysis." This attribute of the 5 V's paradigm is by far the most important from the perspective of data quality. A recent study has shown that an average billion-dollar company in the US is losing about $130 million per year in the US. [11]

Veracity can be further understood as the accuracy of data. Due to the human element as well as the uncertain element which is an inherent property of a system, it is possible to get veracity related issues and these are the most difficult to identify and/or repair. Veracity is most important to make the data operational. The element of bias and uncertainties render the data as less accurate than one expect it for our analysis. It is an established fact that no amount of analysis can replace the need for good and precise data. Terms are reliability, trustworthiness, accuracy, quality, precision, etc… are synonyms of veracity.

E.g. the Englishman would call for a taxi whereas an American would call a cab. Algorithms trained to check one feature may report the data inaccurately. Thus veracity acts as the uncertain factor with the maximum impact. It can also be considered as the factor which differentiates human intelligence from artificial intelligence.

The below chart provides the veracity distribution of Big Data across various domains.

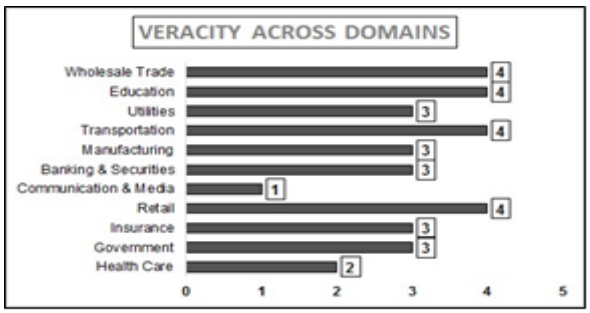


Figure-7: Veracity wise distribution of Big Data across domains. [(Ratings: Very high-5, high-4, avg.-3, low-2, very low-1)]

Given the above study available, the major challenges with veracity are listed below:

- Large volume of generated data and the rate of arrival / transmission of data leads to dilution of original message attributable to noise in the system or inherent causes. [8]
- There is no replacement for human intelligence in the universe and thus the accuracy of 100% can never be achieved by any form of artificial intelligence.
- Since data comes in the form of images, videos and other file types which are laden with large number of data points, storage, retrieval and processing of the data is very time consuming and error prone, not mentioning the costs involved.

VII. ALUE ATTRIBUTE

Understanding the attributes of the data and data analytics leads to asking the million dollar questions. "*Are we building the right system?*" Value attribute is responsible for answering this question. Logical Reasoning says that the destination must be known before embarking on the journey rather than decide along the way or move ahead with a trial and error methodology. In principle one must clearly establish our Critical Expected / Desired Outcome before embarking on the journey of data analytics. Many models, frameworks and methodologies speak about this in different tones. The ITIL framework defines Continual Service Improvement or CSI stage where quantitative process improvement must follow the below 7 steps: [17]

- REQUIREMENT: What is required to be done? This also speaks about the requirements or the business objectives which have been set at the top management level as a direction for the business or the expected outcomes after the further activities are performed.
- RAW DATA REQUIRED: What data is required to conduct the analysis? This is the critical aspect at level two and not at level one as one may be inclined to believe. Working from data to the outcome may end up reaching a destination not desirable.

Also data dump is vast and our objectives are generally crisp requiring only a part of data.

- RAW DATA AVAILABLE: What data is available with us?
- DATA MISSING: What data is missing?
- COLLECTING MISSING DATA: How can one get the data?
- QUALITY OF FINAL DATA: Is the data clean and good?
- TOOLS AND TECHNIQUES: Is/are the methods, tools and techniques for analysis known?

The below chart provides the value distribution of Big Data across various domains.

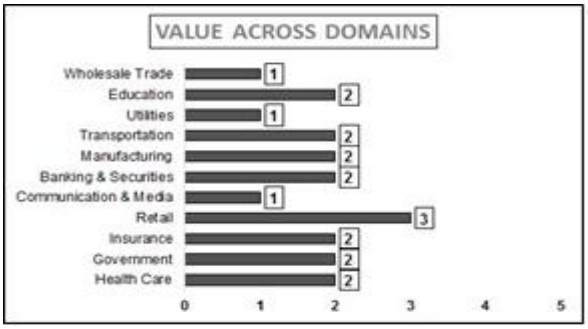


Figure-8: Value wise distribution of Big Data across domains. [(Ratings: Very high-5, high-4, avg.-3, low-2, very low-1)]

Given the above study available, the major challenges with value are listed below:

- Organizations may be biased with their ideas or plans and may miss out vital messages hidden in the data as they would only be looking for something pertinent.

Due to large amounts of data, analysis may take either more time or more money and organizations may be forced to compromise on the level of analysis. This may lead to the skewedness of the results of the analysis. Action plans arising out of the half cooked data may cause more damage than harm.

VIII. 5V'S CORRELATION IN SOME BUSINESS DOMAIN

The earlier section talks about the V Attributes of data and how the domains rank according to the attributes. It is noteworthy to also write about the reverse correlation that exists between the V's within a domain. This can be understood best as a quantitative correlation of percentage contribution of a V-Parameter within a domain towards quality of data attributes. It is expressed as a percentage importance of a V Attribute in the domain using the same data.

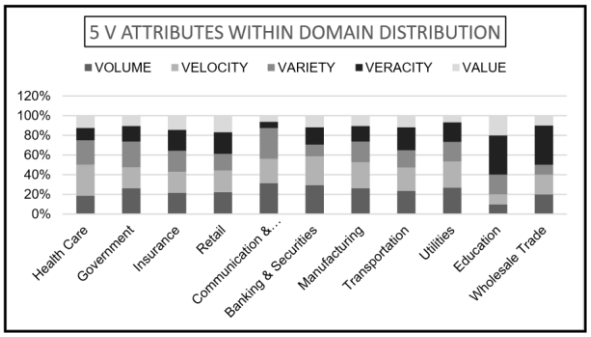


Figure-9: 5V distribution within domain

Below we are now providing the individual domain descriptions.

Wholesale Trade Domain:

Eurostat defines wholesale trade as - "Wholesale trade is a form of trade in which goods are purchased and stored in large quantities and sold, in batches of a designated quantity, to resellers, professional users or groups, but not to final consumers." [12]

The correlation between the V Parameters in this domain is shown below.

0.2 x VOLUME + 0.2 x VELOCITY + 0.1 x VARIETY + 0.4 x VERACITY + 0.1 x VALUE

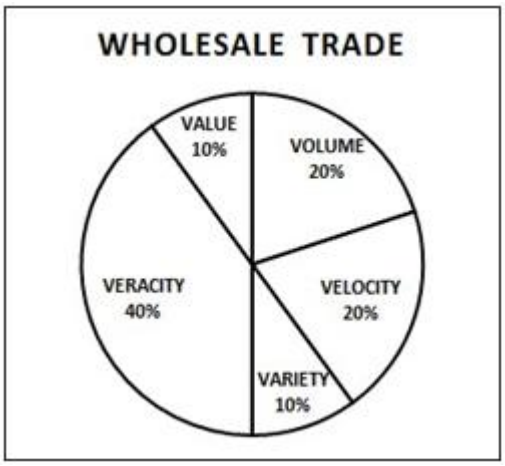


Figure -10: Correlation between the 5 V's in Wholesale Trade Domain Education Domain:

The education domain can be defined as that business segment where learning is imparted at various levels ranging from schools to colleges to Universities to Professional / Corporate Learning and Development programs.

The correlation between the V Parameters in this domain are shown below.

0.1 x VOLUME + 0.1 x VELOCITY + 0.2 x VARIETY + 0.4 x VERACITY + 0.2 x VALUE

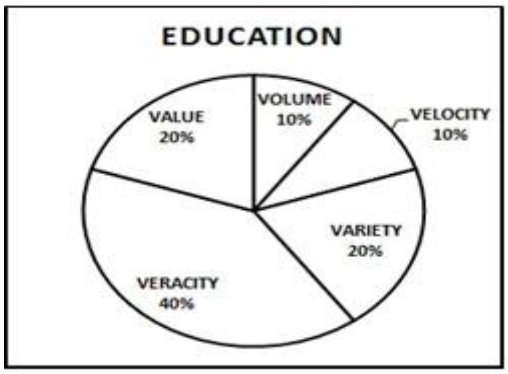


Figure-11: Correlation between the 5 V's in Education Domain Utilities Domain:

The Energy & Utilities Domain Working Group defined the utilities domain as "individuals and organizations engaged in the geospatial aspects of the planning, delivery, operations, reliability and ongoing management of electric, gas, oil and water services throughout the world. [13]
The correlation between the V Parameters in this domain are shown below.

0.29 x VOLUME + 0.29 x VELOCITY + 0.12 x VARIETY + 0.18 x VERACITY + 0.12 x VALUE

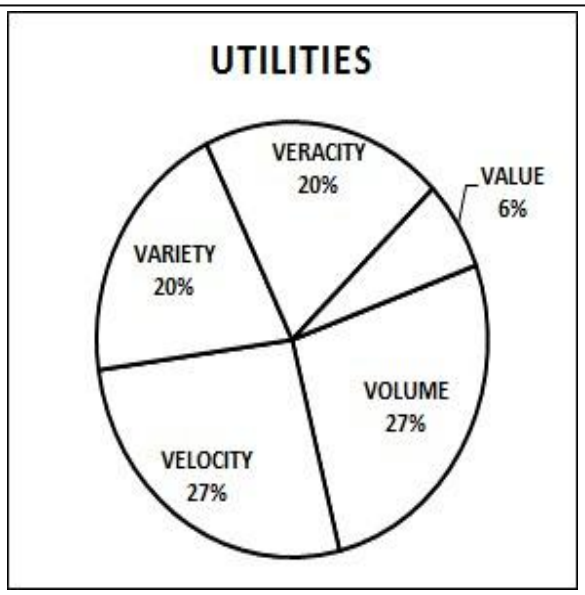


Figure-12: Correlation between the 5 V's in Utilities Domain

Transportation Domain:

The transportation domain can be defined as the business group which is responsible for all travel and transport segments both for human as well as cargo movement. It includes the air, sea and surface transport. The correlation between the V Parameters in this domain are shown below.

0.24 x VOLUME + 0.24 x VELOCITY + 0.18 x VARIETY + 0.24 x VERACITY + 0.12 x VALUE

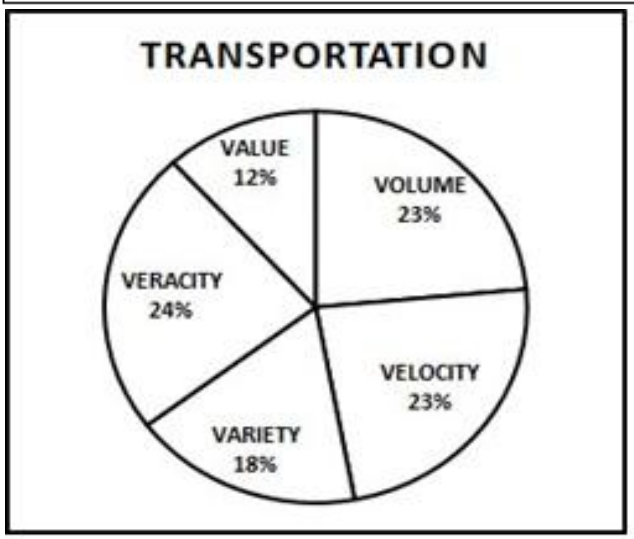


Figure -13: Correlation between the 5 V's in Transportation Domain Manufacturing Domain:

The correlations between the V Parameters in this domain are shown below.

0.26 x VOLUME + 0.26 x VELOCITY + 0.21 x VARIETY + 0.16 x VERACITY + 0.11 x VALUE

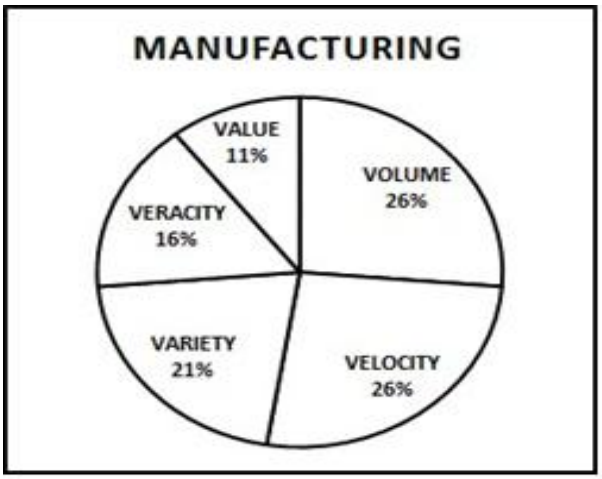


Figure- 14: Correlation between the 5 V's in Manufacturing Domain

Banking and Securities Domain:

Banking and Securities domain as the name suggests deals with Financial Institutions.
The correlation between the V Parameters in this domain is shown below.

0.27 x VOLUME + 0.27 x VELOCITY + 0.2 x VARIETY + 0.2 x VERACITY + 0.07 x VALUE

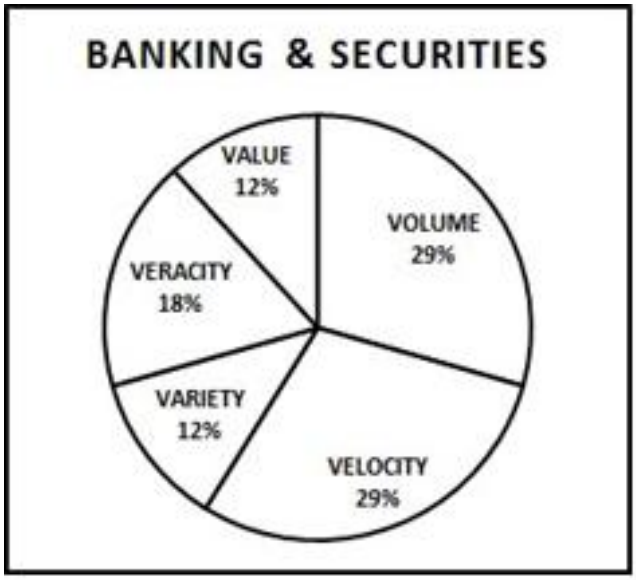


Figure -15: Correlation between the 5 V's in Banking and Securities Domain

Communication and Media Domain:

This domain deals with all types of businesses involving media & communication like print, TV and Movie in addition to the telecommunication sector for mobile and wireless communication. The correlation between the V Parameters in this domain is shown below.

0.31 x VOLUME + 0.25 x VELOCITY + 0.31 x VARIETY + 0.06 x VERACITY + 0.06 x VALUE

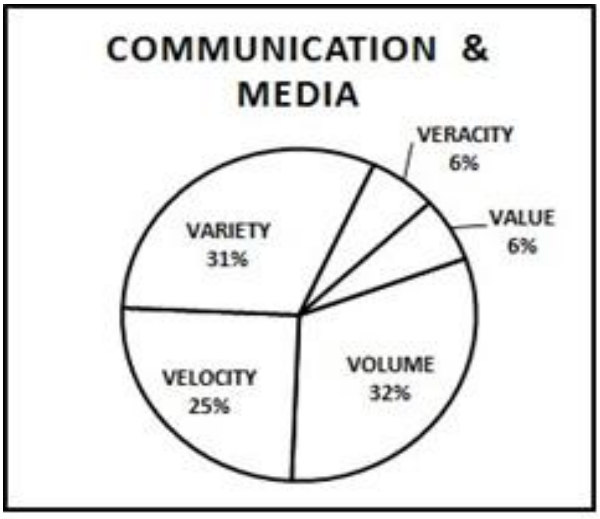


Figure -16: Correlation between the 5 V's in Communication and Media Domain

Retail Domain:

Technofunc defines this domain as "Retail is the sale of goods and services from individuals or businesses to the end-user. The retail industry provides consumers with goods and services for their everyday needs. Retailers are part of an integrated system called the supply-chain." [14]

The correlation between the V Parameters in this domain are shown below.

0.22 x VOLUME + 0.22 x VELOCITY + 0.17 x VARIETY + 0.22 x VERACITY + 0.17 x VALUE

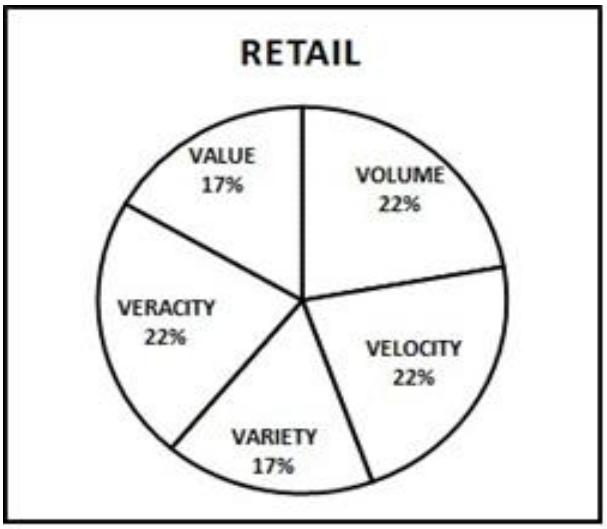


Figure -17: Correlation between the 5 V's in Retail Domain Insurance Domain:

Technofunc defines this domain as "Insurance is a contract between two parties, the insurer or the insurance company and the insured or the person seeking insurance, whereby the insurer agrees to hedge the risk of the insured against some specified future events or losses, in return for a regular payment from the insured as premium." [15]

The correlation between the V Parameters in this domain is shown below.

0.21 x VOLUME + 0.21 x VELOCITY + 0.21 x VARIETY + 0.21 x VERACITY + 0.14 x VALUE

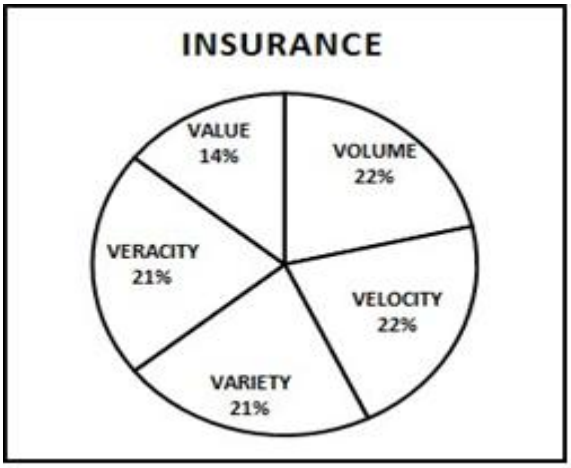


Figure -18: Correlation between the 5 V's in Insurance Domain Government Domain:

With the growth of technology and its availability to the common masses, governments across the world have started migrating to e-Governance models involving remote access and management of government and public interaction.

The correlation between the V Parameters in this domain is shown below.

0.26 x VOLUME + 0.21 x VELOCITY + 0.26 x VARIETY
+
0.16 x VERACITY + 0.11 x VALUE

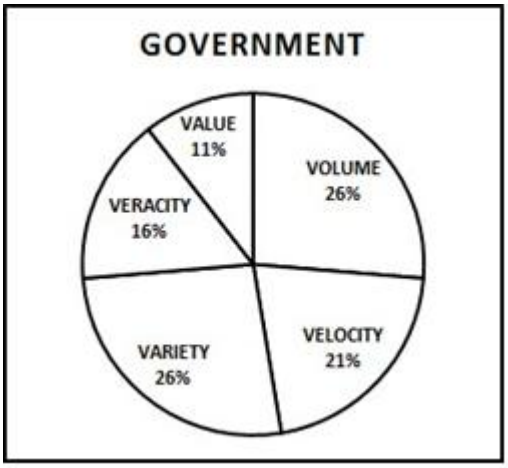


Figure -19: Correlation between the 5 V's in Govt Domain Healthcare Domain:

Wikipedia defines this domain as "Health care or healthcare is the maintenance or improvement of health via the diagnosis, treatment, and prevention of disease, illness, injury, and other physical and mental impairments in human beings. Healthcare systems are organizations established to meet the health needs of target populations." [16]
The correlation between the V Parameters in this domain are shown below.

0.19 x VOLUME + 0.31 x VELOCITY + 0.25 x VARIETY
+
0.13 x VERACITY + 0.13 x VALUE

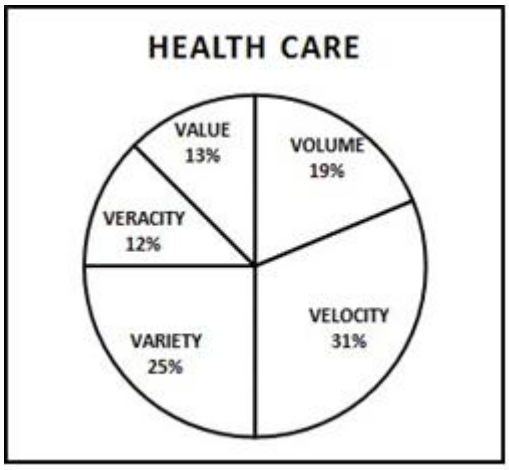


Figure- 20: Correlation between the 5 V's in Healthcare Domain

V. CONCLUSION

The above study reflects and identifies very clear inferences. This survey conclude that data generated in different domains is also distinct in terms of its 5V attributes. That being understood, the challenges within one domain must be analysed and treated unique to the domain rather than a "one-size-fits-all" approach. Comparative study between the domains show us that while roughly a 20% weightage does exist between the 5V's across domains, there is still a subtle but realistic variation that exists. This is a very important finding as it truly shows us that domain has a direct and a very powerful influence on the 5V's or their correlation there off. We would love to believe that domain independence exists when it comes to the importance of the 5V's, but this paper finds adversely. This will pave way for further research and studies to lay importance to the 5V's when designing more sophisticated software of the future in the field of big data analytics. Concluding this we can beyond doubt we can safely say that Big Data clearly deals with issues beyond volume, variety, velocity, veracity and value. Their interdependence is clearly important but that the domain is the key differentiator when it comes to the actual correlation. This is a critical study and one which requires attention by every researcher, academician and person who has an interest in studying systems operating in the space of Big Data and their impacts in various domains.

REFERENCES

1. Bohannon, P., Fan W., Geerts F., Jia X., Kementsietsidis A., () Conditional Functional
2. Dependencies for Data Cleaning, University of Edinburg research publications.
3. Bresina, J L, Morris P H, (2006) Explanations and Recommendations for Temporal Inconsistencies, IWPSS,
4. https://www.stsci.edu/largefiles/iwpss/20066061912 IWPSS_draft4.pdf
5. Brisaboa, Nieves &Luaces, Miguel & Rodriguez, Andrea &Seco, Diego. (2014). An inconsistency measure of spatial data sets with respect to topological constraints. International Journal of Geographical Information Science. 28. 56-82. 10.1080/13658816.2013.811243.
6. Dr. S. Vijayarani and Ms. S. Sharmila, RESEARCH IN BIG DATA – AN OVERVIEW, Informatics
7. Engineering, an International Journal (IEIJ), Vol.4, No.3, September 2016
8. Du Zhang, 'Inconsistencies in Big Data' proceeding, Cognitive Informatics & Cognitive Computing (ICCI*CC), 2013 P. 61-67 12th IEEE Conference.
9. Garboden, Philip. (2020). Sources and Types of Big
10. Data for Macroeconomic Forecasting.
11. 10.1007/978-3-030-31150-6_1.
12. Hartzband, David. (2019). "What Is Data?" DOI: 10.4324/9780429061219-2.
13. Jeffrey Ray, Olayinka Johnny, Marcello Trovati, Stelios Sotiriadis and Nik Bessis, The Rise of Big Data Science: A Survey of Techniques, Methods and Approaches in the Field of Natural Language Processing and Network Theory, Big Data Cogn. Comput. 2018, 2, 22; doi:10.3390/bdcc2030022, www.mdpi.com/journal/bdcc
14. Khan, Samiya& Liu, Xiufeng& Shakil, Kashish&Alam, Mansaf. (2017). A survey on scholarly data: From big data perspective. Information Processing &Management.DOI53. 923-944. 10.1016/j.ipm.2017.03.006.
15. Krogh, Jesper. (2020). Data Types. DOI 10.1007/978-1-4842-5584-1_13.
16. Kumar, Praveen. (2019). BIG DATA ANALYTICS IN HR DOMAIN. DOI 10.1729/Journal.22887.
17. M. V. Martinez, A. Pugliese, G. I. Simari, V. S. Subrahmanian, and H. Prade, How dirty is your relational database? An axiomatic approach, in Proc.
18. 9th European Conference on Symbolic and
19. Quantitative Approaches to Reasoning with Uncertainty, ammamet, Tunisia, LNAI 4724, 2007, pp.103-114.

20. M-C de Marneffe, A. N. Rafferty and C. D. Manning, Finding Contradictions in Text, Proc. of 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2008, pp.1039-1047.
21. Nawsher Khan,Ibrar Yaqoob, Ibrahim AbakerTargio Hashem, Zakira Inayat, Waleed Kamaleldin Mahmoud Ali,1 Muhammad Alam,Muhammad
22. Shiraz,1 and Abdullah Gani, Big Data: Survey, Technologies, Opportunities, and Challenges, Hindawi Publishing Corporation, The Scientific World Journal, Volume 2014, Article ID
23. 712826,https://doi.org/10.1155/2014/712826
24. Özsu, M. &Valduriez, Patrick. (2020). Big Data Processing. 10.1007/978-3-030-26253-2_10.
25. Ptiček, Marina &Vrdoljak, Boris. (2018). Semantic web technologies and big data warehousing. 1214-1219. 10.23919/MIPRO.2018.8400220.
26. Ritter, D. Downey, S. Soderland and O. Etzioni, It's a
27. Contradiction-No, It's Not: A Case Study Using Functional Relations, Proc. of Conference on Empirical Methods in Natural Language Processing, 2008.
28. Samiddha Mukherjee, Ravi Shaw, Big Data – Concepts, Applications, Challenge and Future Scope, International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 2, February 2016, ISSN (Online) 2278 – 1021, ISSN (Print) 2319 – 5940
29. Sergio Luján-Mora, Manuel Palomar, 'Reducing Inconsistency in Integrating Data from Different Sources'. Proceedings 2001 International Database Engineering and Applications Symposium (IDEAS
30. 2001), p. 209-218: IEEE Computer Society, Grenoble (France), July 16-18 2001.
31. https://doi.org/10.1109/IDEAS.2001.938087
32. Smirnov, Alexander &Levashova, Tatiana &Shilov, Nikolay. (2012). Ontology Alignment for IT Integration in Business Domains. 127. 153-164. 10.1007/978-3-642-34228-8_15.
33. Yaqoob, Ibrar& Hashem, Ibrahim &Gani, Abdullah & Mokhtar, Salimah& Ahmed, Ejaz &Anuar, Nor &Vasilakos, Athanasios. (2016). Big Data: From Beginning to Future. International Journal of Information Management. 36. 10.1016/j.ijinfomgt.2016.07.009.
34. Zhang, On Temporal Properties of Knowledge Base Inconsistency. Springer Transactions on Computational Science V, LNCS 5540, 2009, pp.20-37.
35. https://pediaa.com/what-is-the-difference-between-d ata-redundancy-and-data-inconsistency/
36. https://techcrunch.com/2012/08/22/how-big-is-faceb ooks-data-2-5-billion-pieces-of-content-and-500-ter abytes-ingested-every-day/
37. https://www.bizdata.com.au/blogpost.php?p=costs-o f-data-redundancy-and-data-inconsistency
38. https://www.brainkart.com/article/Relationship-Typ es,-Relationship-Sets,-Roles,-and-Structural-Constra
39. ints_11431/#:~:text=This%20constraint%20specifie s%20the%20minimum,called%20the%20minimum %20cardinality%20constraint.&text=We%20will%2 0refer%20to%20the,constraints%20of%20a%20rela tionship%20type.
40. https://www.dsayce.com/social-media/tweets-day/
41. https://www.emc.com/leadership/digital-universe/20 14iview/executive-summary.htm
42. https://www.happiestminds.com/Insights/big-data-a nalytics/
43. https://www.heshmore.com/how-much-data-does-go ogle-handle/
44. https://www.kdnuggets.com/2012/12/idc-digital-uni verse-2020.html
45. https://www.techopedia.com/definition/19504/functi onal-dependency#:~:text=Functional%20dependenc y%20is%20a%20relationship,is%20functionally%2 0dependent%20on%20X.
46. https://www.washingtonpost.com/national/health-sci ence/google-says-one-hour-of-video-is-now-being-u ploaded-to-youtube-every-second/2012/01/27/gIQA tubBdQ_story.html

AUTHORS' PROFILE

Vinaya Keskar is a Research Student in Computer Science Department Savitribai Phule Pune University (SPPU) She received her Masters' degree from North Maharashtra University, Jalgaon MS, India. Her research interests are Algorithms, System programming, Theory of computation, Big Data, and Teaching in addition to being an ardent and experienced academician.

Dr. Jyoti Y. Yadavis an Assistant Professor at Department of Computer Science, Savitribai Phule Pune University. She is also a research guide assisting research scholars in fulfilling their academic aspirations. Her research interests are Cloud Computing, Cloud Based Systems, Google Cloud Solutions, Computer Programming in JAVA, SQL and C++ predominantly and Computer Algorithms.

Dr. Ajay H. Kumar is a Research Guide in Computer Science Department, Savitribai Phule Pune University and Director of Jaywant Technical Campus. His research interests are Computer Networks, Data Warehousing, Cloud Computing and performance of communication systems.