ISSN NO.:2350-1146 I.F-5.71

V's as a basis of Big Data & Data Intensive Science Discoveries

Anshuman Dwivedi
Birla Institute of Technology - Pilani
Hyderabad Campus
Hyderabad, India
anshuman.dwivedi@gmail.com

Dr. Vanita Joshi
ICFAI Business School
Mumbai, India
Vanita207@gmail.com

***Abstract*—This survey attempts to consolidate the hitherto fragmented discussions on big data and its harness potential to extract the knowledge. Firstly, various definitions and the features of Big Data are analyzed. Secondly, we have surveyed the several existing and new aspects of the data-intensive scientific discovery in terms of Vs and concluded that the systematic treatment of Vs can convert "data-centric organizational" into "knowledge-centric organizational". At last, based on the various research papers available, we have derived a probable big data dimensions' model as 6V-6O.**

Keywords — Big Data; Data-Intensive Scientific Discovery; Volume; Velocity; Variety; Veracity; Value; Viability; Validity; Volatility; Variability; Visualisation

I. INTRODUCTION

In 2008 Google launched the project "Google Flue Trends"- a google-search based model to predict flu activities. With this Google was able to "accurately estimate the current level of weekly influenza activity in each region of the United States with lag of just one day". [1] But later this converted into "false" because Google's results were merely based on the statistical patterns of the data grounded on correlations of some specific search queries of spreading of influenza in certain space and time without any causal analysis. In summary, they cared about the "correlations rather than causations". This is one of the common problem of large and complex set of data. It shows that traditional analysis of data cannot solve problems alone but big insights are required to solve the issues. Recently, Data-Intensive Scientific Discovery (DISD) [2] along with "Big Data" has emerged as the fourth scientific paradigm after "empirical science", "theoretical science" and "computational science" [3] in order to solve the larger and complex data related issues.

According to researchers, "Big data has arrived, but big insights are yet to come" [4]. The challenge to accurately solve the problems is yet to be discovered. In order to avoid misconception and to characterize the large and complex set of data properly, a set of attributes were identified as major characteristics. This set of attributes in scientific world is referred to as the V's of Big Data, as per their initial names. [5] In this paper, we have surveyed the several aspects of the data-intensive scientific discovery (DISD) in terms of Vs and concluded that the systematic treatment of Vs can be very helpful to "analyze" and convert complex data into "big opportunities" and also has the potential to replace the statistical practice of sampling to estimate any phenomena by analysis of full populations.

In Section II, we have discussed several motivational factors to write this paper. The opportunities and challenges aroused from traditional Vs are introduced in Section III to V. Then, we have given a detailed demonstration of new state-of-the-art Vs to handle data-intensive applications in Section VI to XIII. At last - XIV, we have researched the various papers available on ieeexplore and have derived a probable big data dimensions' model as 6V-6O. In the XV and XVI, we have drawn the 'areas of concern & future work' and conclusion.

II. MOTIVATIONS

Despite big data's popularity, it has yet to obtain a concrete and universally accepted definition [6]. Probably it's because "big data" is not the result of a specific innovation but rather is a blend of revolutionary and evolutionary changes happened both in technical and research arena [7]. At the moment, 'big data' term is primarily used as an umbrella term to "define data sets that are large, complex and beyond the traditional computational processing limits" [8].

The term "big data" was first coined by david ellsworth and michael cox of NASA's Ames Research Centre in the Proceedings of the IEEE 8th conference on Visualization. In this paper they mentioned - "Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data. When data sets do not fit in main memory (in core), or when they do not fit even on local disk, the most common solution is to acquire more resources." According to ACM digital library this was the first paper to use the term "Big Data" [9] but some authors also give credit to John Mashey for making this term popular.[10, 11]

The term Big Data is so generic that to give credit to those who used this word combination early is not a good idea. Instead, we should recognize those who first used this to represent lot of data along with new ways to handle it.[10]

Professionally, Gartner was the first firm within the industry to name it as "big data" in 2001. According to them we can define the challenges and opportunities in data growth as factor of three 'Vs': volume, velocity and variety [12]. As per the Gartner researchers –

"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."

The McKinsey Global Institute also published a report in 2011, according to them the most subjective definition of Big Data is –

"Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze."[13]

This definition is quite general and captures only the size aspect of the available data and demonstrates the need of technology to handle big data sets properly. This definition does not clearly defines the growing of data size; in fact, it emphasizes the only technology need. Likewise, Devise and Patterson also points it out as "the data that is too big to be handled and analyzed by traditional database protocols such as SQL"; [14] But Edd Dumbill explicitly recognizes it as multi-dimensional factor. According to him, "the big data is too big, moves too fast, or doesn't fit the strictures of your database architectures". [6] Clearly, the extra characteristics are always required to considered complex data as Big Data.

The most popular and agreed definition of Big Data is the above definition provided by Gartner. It describes the data set as Big Data if it is formidable to perform capture, curation, analyze and visualize it using available technologies.[15] Volume refers to the increased amount of data being generated from different sources, velocity refers to the pace of data generation and interaction, and variety refers to incompatible data formats, structures, and semantics.

Apart from above three Vs, later studies [12, 13, 6,17,16,14,18,9] pointed out that the traditional definition of 3Vs is insufficient to explain the big data we face now. Many researchers and institutes like IEEE focus on veracity, value, viability, validity, volatility, variability and visualization [15] to face the new challenges. "Variability" and "Value" denote the diversified nature and significance of data set."Visualization" is added by some researchers to represent hidden knowledge more intuitively and effectively by using different graphs. "Veracity" is also used to describe truthfulness of the data. A recent articles also shows that existing definitions of big data also requires the way it creates value for organizations.[13]

In industry, Oracle treats big data as the natural evolution from traditional relational database driven by business decision makings and surrounded by the sources of unstructured or semi structured data. In real sense, the Oracle definition is not very clear about exactly when the term big data should be used? Or is this just another name of schema-on-read approach? [16] Intel defines big data in the quantified way, rather than providing it in a qualitative way. They treat big data as "generating a median of 300 terabytes (TB) of data weekly" [17]

As per Microsoft, "Big data is the term increasingly used to describe the process of applying serious computing power - the latest in machine learning and artificial intelligence - to seriously massive and of- ten highly complex sets of information" [18]. Google also provides the Big Data defining as to technologies and initiatives that involve data that is too

Finally, we can refer big data as a combination of informing decision making with analytical insight using a set of enabling technologies at its most granular level for very large and diverse sources of data.

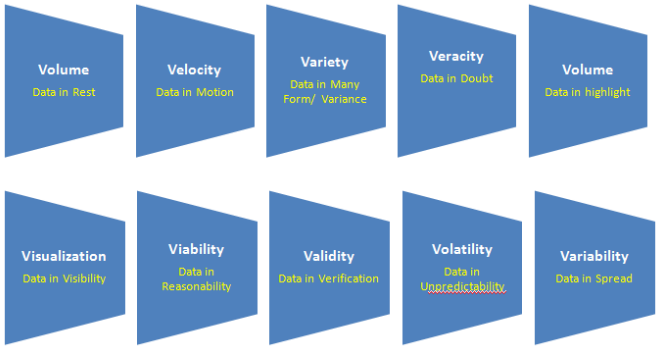

Fig 1: Big Data : In context of V

III. 1ST V – VOLUME: DATA IN REST

Volume of Big data refers to the size of data being generated from different sources. This aspect immediately comes to our mind when we think about Big Data.

The ability to process large amount of data has been the central attraction of Data Analytics. According to Moore's law, the CPU performance i.e. the number of transistors that can be placed on an integrated circuit is increasing exponentially, virtually doubling in each 18 months and though not directly connected to Moore's Law, but still disks are also following the semiconductors almost at the same rate. However, the disks' rotational speed has not improved at the similar rate (also referred as Kryder's Law") [20]. This can be interpreted as the information is increasing at exponential rate [21], but the information processing methods are relatively slower. This can be concluded that in real world the state-of-the-art techniques and technologies are not capable to solve the growing data problems on real-time. Another important question is: what is responsible for the data growth? As per researchers, the main drivers of the data explosion are – switch from analog to digital [21], rapid increase in data generation by social media and "smart" devices. Further, bringing code to larger volume of data is also a bandwidth intensive task.

The study done by information-management organization - EMC and IDC in 2007 showed that the amount of digital information created and replicated in the same year outdid the world's data storage capacity for the first time. [22]. According to Forbes the data been created in the past two years is more than the entire history of the human race and about 1.7 megabytes of new information is being created every second by every individual on the planet.[23] and as-of-now 2.8 Zetabytes of data exist in the digital universe today. Today, there are more than 2.07 billion users in Facebook, and the overall size of semi or unstructured generated data is 30+ Petabytes. Out of which 66.1% users' login to every month and send on an average 31.25 million messages to each other. Similarly, Walmart also handles more than 1 million customer transactions every hour, which is more than 2.5 petabytes of structured data.

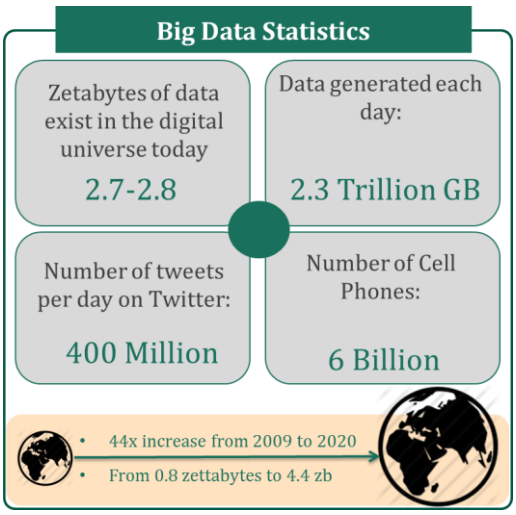

Fig 2: Data Growth

Concluding, "The more data that is created, the more knowledge […] people can obtain" [24] provided they have big data technologies.

IV. 2ND V – VELOCITY: DATA IN MOTION

Another most important aspect related to big data is the velocity or the pace at which data is being generated and analyzed.

One of the commercial advertisements from IBM points out that "*you wouldn't cross the road if all you had was a five-minute old snapshot of traffic location.*" There are situations where one has to be prepared with technology and concepts to process the data immediately rather wait for some program to generate report or some job to complete. [25] This is advantageous to business users too to make valuable decisions on time and get strategic advantages and fast ROI. According to some researchers like Teece, "The postware evolution of markets has powerful strategic implications […] It is no longer in product markets but in intangibles assets where advantage is built and defended." [26]

Velocity is usually considered in terms of volume of data and but its internally its related to processing time at which it is created and derive knowledge or insights. [27] The two popular approaches of data analytics are "analyze and store approach" and the "store and analyze" approach. The first approach is used in traditional BI and the second "works well when implementing business process management and service oriented technologies" [28] and is particularly useful for monitoring and analyzing real time activities.

Time reduction is also another common objective of big data technologies. Generating hundreds of thousands of models on data set and to aggregate them immediately is the key differentiator between traditional and high performance computing.

To store, analyze, and process a fast-moving dataset, open-source document database and leading NoSQL databases such as MongoDB, HBase and Cassandra are used. For example, Facebook uses HBase, to handle 350 million users with over 15 billion messages per month. Standard relational databases can't easily deliver the same performance. [29]

Another term "Viscosity" is also used to demonstrate the latency or lag time characteristic of the data but as per some researches this can be understood as an element of Velocity.

Concluding, big data can also be characterized by its velocity: the pace at which the data it is being generated and insight or knowledge is extracted from it.

V. 3RD V – VARIETY: DATA IN MANY FORMS/VARIANCE

Variety is the most significant problem of big data. According to Boyd & Crawford, big data is not notable because of its size, but its because of its multiplicity and relationship with other data.[30] It also denotes the amount of data that can be trusted. Today most of the organizational data are dark by nature because they are less used or unused, similar to dark matter in physics.

Big data comes in all kind of forms – structured(data that is rrelational in nature), semi-structured (web logs, social media feeds, raw feed directly from a sensor source, email, etc.) and un-structured(video, still images, audio, clicks). Analyzing these data can lead to potential value and exactly for this reason, big data analytics differs from traditional data analytics in terms of volume and velocity.

The Traditional analytics is not able to handle the data that beyond the capability of relational or hierarchical data base engines. Over 80% of the data generated is unstructured, which traditional tools cant handle and analyze.

There are two aspects of variety : syntax and semantics. Traditionally, these two parameters have been a benchmark to determine the reliability of data. But modern ETL tools are very capable of dealing the data in as is format, virtually in any syntax. In the past they were very less able to deal with semantically rich data such as free text. A good example of this is the linking of a customer database with social media

data sources such as Twitter. This "linked data" often uses standardized ways of data transportation using HTTP, preserving context making it both usable for humans and computers.

Concluding, "Data variety is a measure of the richness of the data representation. From an analytic perspective, it is the biggest obstacle and opportunity to effectively process incompatible data formats, non-aligned data structures, and inconsistent data semantics".

VI. 4TH V – VERACITY: DATA IN DOUBT

One of the new identified challenge when dealing with big data is veracity or the trustworthiness of data. Technically, this can be considered as the quality of the data (i.e. correctness, accuracy etc...)

Veracity is the hardest thing to achieve with big data due to its inherent nature of high volume and variety. The biggest problem is if the data is inaccurate, redundant or unreliable, the overall purpose of analysis can lead to incorrect results and all the information can lead to a useless and very expensive Big Data environment.

According to the Roberto, "There are several challenges: How can we cope with uncertainty, imprecision, missing values, misstatements or untruths? How good is the data? How broad is the coverage? How fine is the sampling resolution? How timely are the readings? How well understood are the sampling biases? Is there data available, at all?"[31]

The alarm related to the veracity apply both to the input as well as to the result. Malfunctioning IoT sensors, incorrect social media feeds, heterogeneity of measuring devices or logs, all need to be accounted for during the ETL phase. Handling accuracy through volume is also not a good idea always because it requires a lot of extra computing power.

Concluding, Veracity in is the measure of the trustworthiness of data, which can't be compromised always.

VII. 5TH V – VALUE/ VOLATILITY: DATA IN HIGHLIGHT

The ultimate goal of the big data is to fabricate some value. Dealing with big data should be an efficient trade-off between investment on performance and result. In turn this directly depends upon the governance mechanism i.e. the available policies and structures that we eventually bring balance between reward and risk of the data [32]. Same time, these policies and structures, if not carefully written and implemented may restrict businesses to extract true value of data.

Value comes only from what we infer from it. According to Werner Vogels, CTO of Amazon.com - "in the old world of data analysis you knew exactly which questions you wanted to asked, which drove a very predictable collection and storage model. In the new world of data analysis your questions are going to evolve and changeover time and as such you need to be able to collect, store and analyze data without being constrained by resources." So in true sense, the model of big data gives value to the business.[32]

According to MGI, the "value" comes by – "Creating transparencies; Discovering needs, exposing variability, and improving performance; Segmenting customers; and Replacing/supporting human decision-making with automated algorithms". [13]

In summary, there is no single set formula for extracting value from Big Data; It entirely depends on the data, the algorithm and governance model.

VIII. 6TH V – VISUALISATION

Data visualization is the direct process to interact with data i.e. going beyond the traditional analysis and giving presentation to it. Another most referred definition found in research community is "the use of computer-supported, interactive, visual representation techniques of data to amplify cognition" [33] [34]. The main goal of Data Visualization is to explore the content of the data, identify sense-making patterns, correlations and causalities. It is the process that transforms symbolic representation to geometric representation. One of the most important strength of visual analytics is to engage the whole of our perceptual and cognitive capabilities to the analytical process.

The Big Data visualization is growing rapidly. As per Mordor Intelligence [35] the visualization market will increase at 9.21 % from $4.12 billion in 2014 to $6.40 billion by the end of 2019. The research conducted by International Data Group (IDG) also shows that 98 % of the most effective companies working with Big Data are using visualization as main analysis, presentation and decision-making purposes e.g. Amazon, Twitter, Apple, Facebook and Google [36, 37].

In a business environment, visualizations refers two broad goals - Explanatory and Exploratory. In Big Data context, exploration and visualization has many challenges [38,39,40,41, 42] probably it's because of the complexity in 3Vs or 4Vs.

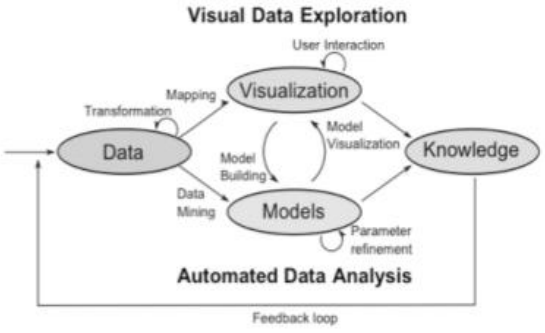

Fig 3: The visual analytics process, adapted from [43]. It shows the interaction between the user, the data and the visual representations.

First major challenge of big data visualization is to conduct data visualization on larger, diversified and heterogeneous

data sets. In this respect, the visualization and exploration systems must have enough computing power to handle the data and equipped with efficient query and algorithms to handle the complex analysis.

In a "big data visualization" context the traditional database-oriented systems cannot be adopted. They have poor performances in functionalities, scalability and response time too. In order to handle these issues, a big data systems usually adopts data reduction or approximation techniques based on: (1) sampling and filtering [44,45,46,47,48]; or/and (2) aggregation (e.g., binning, clustering) [48, 49, 50, 51, 52, 53] (3) query-based approaches like query translation, query rewriting etc. [53, 54] (4) incremental (a.k.a. progressive) techniques based on user interaction or time (over progressively larger samples of the data) [34, 53, 54] (5) adaptive indexing approach where indexes are created incrementally and adaptively throughout exploration. (6) Some other approaches, related to parallel architectures as mentioned inhttp://graphics.cs.wisc.edu/Vis/CompIV/

Perceptual and uncertainty are also major challenges of big data visualization. New frameworks for modelling, handling and characterizing the uncertainty are highly necessary now for the analytical processes. Visualization at the micro level unusually leads to over-plotting and in turn affect the perceptual and cognitive capacities; dropping the data through sampling-filtering or aggregation can also ground the removal of interesting structures or outliers.

To sum up, more operative and innovative approaches are required that can visualize the large number of data objects, using a limited number of resources.

IX. 7TH V – VIABILITY

Viability means the quality variable to maintain itself or having a reasonable chance of success. [55, 56] In other words, this is the process to validate that hypothesis along with variables have meaningful impact on desired or observed outcomes.

X. 8TH V – VALIDITY: DATA VERIFICATION

Verification of results is important. "With big data, we must be extra vigilant with regard to validity. For example, in healthcare, you may have data from a clinical trial that could be related to a patient's disease symptoms. But a physician treating that person cannot simply take the clinical trial results as without validating them." [57]

XI. 9TH V – VOLATILITY

In big data world, data retention period is very much significant word than traditional data world. Excess of Big data retention period may cause extra cost, storage and security. [59]

XII. 10TH V – VARIABILITY

Rationally, Variability is the measure of the spread. It is the extent to which data points in a statistical distribution are spread from the mean or differ from each other.

Variability can be associated with any of the V, depending upon the context. For example, in some situations - volume is high, velocity is low but the associated value can be high. Whereas, on some occasions, the velocity and variety are low but the value is significant.

Variability is also used to represent the statistical problem of 'outlier'.[60]

XIII. OTHER: 3C - COMPLEXITY, COST AND CONSISTENCY, ACCESSIBILITY, QUALITY, VELOCITY

- 3C sub-dimension - complexity, cost and consistency are also mentioned by some authors.[61]
- Accessibility and Quality - are also used by Morabito as dimensions [62, 63]

XIV. DIMENSIONS MODEL : 6V-6O

In this section, we have analyzed all dimensions available in various big data literatures.

TABLE I. DIMENSIONAL ANALYSIS

Dimension	Authors	Frequency
Volume	Gartner, IDC etc…	>= 200
Velocity	Gartner, IDC etc…	>= 200
Variety	Gartner, IDC etc…	>= 200
Veracity	Buhl et al. (2013); Morabito (2014); Ali-ud-din Khan, Uddin, Gupta(2014); Bedi, Jindal, Gautam (2014); Hansmann and Niemeyer (2014); Ebner, Bühnen, Urbach (2014); Roberto(2015), Osden Jokonya(2015), Daniel E. O'Leary(2015) etc…	>= 20
Value	IDC etc…	>= 200
Viability	Bedi, Jindal, Gautam (2014); Neil, Biehn (2013)	<=10
Validity	Ali-ud-din Khan, Uddin, Gupta (2014), O. C. Ubadike(2015)	<=10
Volatility	Ali-ud-din Khan, Uddin, Gupta (2014)	<=10
Accessibility	Morabito (2014)	<=2
Quality	Morabito (2014)	<=2
Variability	Katal, Wazid, Goudar (2013) Bedi, Jindal, Gautam (2014)	<=10
Complexity	Katal, Wazid, Goudar (2013)	<=2

Visualisation	Anupama RamanDhivya (2015) etc..	>= 20

In Table 1, (Appendix-I) mentions the frequency count per dimension along with some researchers. During the literature review, we have observed that Gartner's 3 V's model - volume, velocity and variety and IDC's added dimension - value are the most commonly referred dimensions. However, later studies [6,9,12,13,16,17,18] denote that these traditional definitions are insufficient to explain the upcoming challenges related to big data. Many researchers and institutes like IEEE are also in process of recognising the importance of veracity and value to deal new challenges. [15] "Visualization" is also considered by some researchers to represent hidden knowledge more intuitively and effectively. "Veracity" is also used to describe truthfulness of the data. At last, as per our observation in big data related research 6V's – Volume, Variety, Velocity, Value, Visualisation and Veracity can be referred as core dimensions and other 6 - Viability, Validity, Volatility, Variability, Cost and Complexity can be picked up as extra-dimensions.

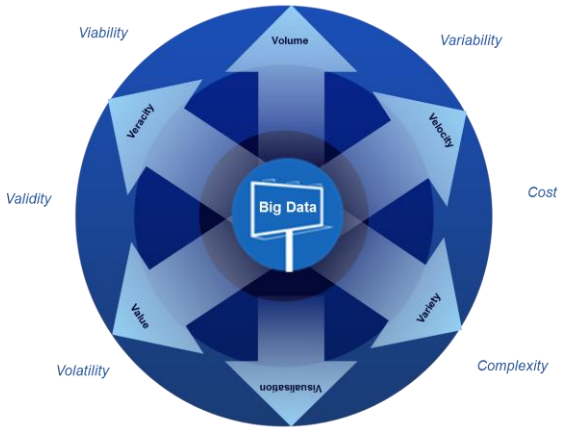

Fig 4: 6V-6O Models

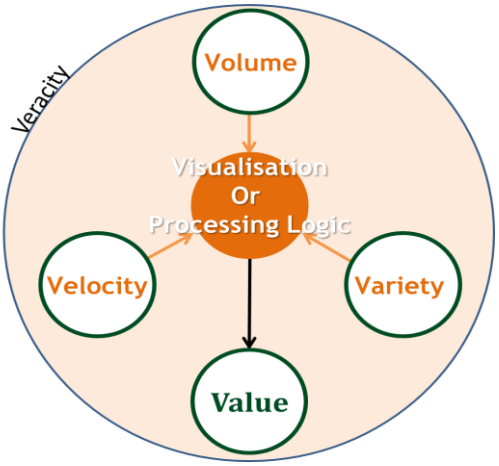

Fig 5: V's of 6V-6O Models

XV. Areas of Concern & Future Work

After studying various researches and current state of art approaches, we found that -

1. The processing capabilities are almost getting doubled in every 18 months. However, the data access rate i.e. disks' rotational speed is still slow.
2. Information explosion is at exponential rate, but the information processing algorithms are still relatively slow.
3. The new big data technologies are not able to solve real-time analysis.
4. Developing algorithms to utilize all major Vs is still a challenge.
5. Big Data related inconsistence and incompleteness, scalability, timeliness and data security are major issues.
6. Cost-effective devices for large-scale data is another hot topic.
7. Visualisation for big data is also an area that is attracting attention of researchers and practitioners.

XVI. Conclusion

As the digital work is expanding exponentially, we have entered into an era of next frontier of data science 'to provide the centralized highly scalable and cost effective modelling platforms to discover new patterns using data driven models'. There is no doubt that big data also means big systems, big challenges, big opportunities and lot of research work is required to understand it properly.[19]

In this paper, we have surveyed various Vs of big data in order to understand various dimensions conceptually. Finally, we conclude by –

"The prime objective of big data is to provide the centralized highly scalable and cost effective modelling

platforms to discover new patterns using data driven models. The survey from various researchers demonstrates that in terms of challenge it's not just about 3 or 5 V's but it's almost about '6 primaries and 6 other' Vs.

...And systematic treatment of Vs can convert 'organizational dark data' into 'big opportunities'".

REFERENCES

[1] Jeremy Ginsberg1 , Matthew H. Mohebbi1 , Rajan S. Patel1 , Lynnette Brammer2 , Mark S. Smolinski1, and Larry Brilliant, "Detecting influenza epidemics using search engine query data", Nature, February 2009, See: doi:10.1038/nature07634.

[2] Gordon Bell, Tony Hey, Alex Szalay, "Beyond the data deluge", Science, 2009.

[3] Tony Hey, Stewart Tansley, Kristin Tolle, "The fourth paradigm: data-intensive scientific discovery", Microsoft Research, 2009.

[4] Tim Harford, "Big data: are we making a big mistake?", Finiancial Times, March 28-2014, See: https://next.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0.

[5] Radu Tudoran., "High-Performance Big Data Management Across Cloud Data Centers", Phd Thesis, Radu-Marius Tudoran (ENS Rennes - IRISA / KerData), Computer Science, page 24, Jan 2015.

[6] Edd Dumbill, "Defining Big Data". May 2014, Forbes, http://www.forbes.com/sites/edddumbill/2014/05/07/defining-big-data/#717eecb814d0.

[7] J. Gantz, "Extracting value from chaos", IDC research report IDC research report, page 6, 2011.

[8] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A. Byers, "Big data: The next frontier for innovation, competition, and productivity", 2011.

[9] Michael Cox, David Ellsworth, "Application-controlled demand paging for out-of-core visualization", Proceedings of the IEEE 8th conference on Visualization, 1997.

[10] Gil Press, A Very Short History Of Big Data, Forbes, May 2013, http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#4da85c8a55da

[11] John R. Mashey, "Big data ... And the next wave of infrastress", Slides from invited talk. Usenix, 25 April 1998.

[12] Pettey & Goasduff, "The three Vs", Gartner, 2011.

[13] Sulayman K. Sowe, Koji Zettsu, Curating, "Big data made simple: perspectives from scientific communities", Big Data. Mar 2014.

[14] K. Davis, D. Patterson, "Ethics of big data: balancing risk and innovation", O'Reilly Media, 2012.

[15] Qiang Yang, "Introduction to the IEEE Transactions on Big Data", Jan 2015.

[16] Oracle White Paper, "Information management and big data : a reference architecture", September 2014, http://www.oracle.com/technetwork/database/bigdata-appliance/overview/bigdatarefarchitecture-2297765.pdf.

[17] Peer Research Big Data Analytics Intel's IT Manager Survey, "How organizations are using big data", August 2012, http://www.intel.co.za/content/dam/www/public/us/en/documents/reports/data-insights-peer-research-report.pdf.

[18] Microsoft News Center, "The big bang: how the big data explosion is changing the world", Feb 2013, https://news.microsoft.com/2013/02/11/the-big-bang-how-the-big-data-explosion-is-changing-the-world/#sm.01xetidn151gcoi11vs1n221chrbm.

[19] Hopkins and Evelson, "Big opportunities in Big Data", 2012.

[20] Microsoft Research Technical Report, "Spending Moore's Dividend", MSR-TR-2008-69.

[21] "The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things", EMC Digital Universe with Research & Analysis by IDC, April 2014, http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm.

[22] "Big Data: 20 Mind-Boggling Facts Everyone Must Read", http://wikibon.org/blog/big-data-statistics/.

[23] Dave Evans, "The internet of things how the next evolution of the internet is changing everything", Cisco, April 2011.

[24] Bhuvan Unhelkar, Big Data Strategies for Agile Business, CRC Press, 2017

[25] David J Teece, "Dynamic capabilities and strategic management", Strategic Management Journal,1997.

[26] David J Teece, Essays in Technology Management and Policy: Selected Papers of David J. Teece, World Scientific, 2003. http://www.haas.berkeley.edu/faculty/pdf/teece_david.pdf.

[27] Pettey & Goasduf, "gartner says solving 'big data' challenge involves more than just managing volumes of data", June 2011.

[28] C. White, "Using big data for smarter decision making", BI research, 2011.

[29] Benoy Bhagattjee, "Emergence and taxonomy of big data as a service", May 2004, http://web.mit.edu/smadnick/www/wp/2014-06.pdf.

[30] Danah Boyd, Kate Crawford , "A decade in internet time: symposium on the dynamics of the internet and society", September 2011.

[31] Roberto V. Zicari, "Big Data: Challenges and Opportunities", Goethe University Frankfurt, Oct 2015.

[32] Dr. Werner Vogels, CTO Amazon, www.kdnuggets.com/2011/11/zicari-interview-**amazon-cto-werner-vogels**.html

[33] S. Kard, J. MacKinlay, and B. Shneiderman, "Readings in information visualization: using vision to think", Morgan Kaufmann, 1998

[34] Alfredo R. Teyseyre and Marcelo R. Campo, "An overview of 3d software visualization", IEEE Transactions on Visualization and Computer Graphics, vol.15, No.1., 2009.

[35] Report, "Data visualization applications market future of decision making trends, forecasts and the challengers", Mordor Intelligence; 2014.

[36] SAS, "Data visualization: making big data approachable and valuable", Market Pulse: White Paper, 2013.

[37] P. Simon, "The visual organization: data visualization, Big Data, and the quest for better decisions", John Wiley & Sons, 2014.

[38] Nikos Bikakis, Timos Sellis, Workshop Proceedings of the EDBT/ICDT 2016 Joint Conference Exploration and Visualization in the Web of Big Linked Data: A Survey of the State of the Art, 2016.

[39] N. Bikakis and G. Papastefanatos, "Visual exploration and analytics of big data: challenges and approaches", 2016.

[40] J. Heer, S. Kandel, "Interactive analysis of big data", ACM Crossroads, 2012.

[41] C. Tominski, J. Abello, and H. Schumann, "CGV - an interactive graph visualization system", Computers & Graphics, 2009.

[42] E. Wu, L. Battle, S. R. Madden, "The case for data visualization management systems", 2014.

[43] D. Aurelio, "Visualizing information associated with architectural design variations and simulations", HCI International, July 2013.

[44] D. Fisher, I. O. Popov, S. M. Drucker, and M. C. Schraefel, "trust me, i'm partially right: incremental visualization lets analysts explore large datasets faster", Microsoft Research, May 2012.

[45] Y. Park, M. J. Cafarella, and B. Mozafari, "Visualization-aware sampling for very large databases", ICDE, 2016, https://arxiv.org/pdf/1510.03921.pdf.

[46] S. Agarwal, B. Mozafari, A. Panda, H. Milner, S. Madden, and I. Stoica. "BlinkDB: queries with bounded errors and bounded response times on very large data" In EuroSys, 2013.

[47] J. Im, F. G. Villegas, and M. J. McGuffin, "Visreduce: fast and responsive incremental information visualization of large datasets. in bigdata", 2013. [8-42, 8-25, 8-74, 8-73, 8-97, 8-138, 8-96, 8-1, 8-15, 8-71].

[48] N. Elmqvist and J. Fekete, "Hierarchical aggregation for information visualization: overview, techniques, and design guidelines", TVCG, 16(3), 2010.

[49] N. Bikakis, G. Papastefanatos, M. Skourla, and T. Sellis, "A hierarchical aggregation framework for efficient multilevel visual exploration and analysis", 2015, http://arxiv.org/abs/1511.04750.

[50] U. Jugel, Z. Jerzak, G. Hackenbroich, and V. Markl, "VDDA: automatic visualization-driven data aggregation in relational databases", VLDBJ, 2015.

[51] Z. Liu, B. Jiang, and J. Heer., "Real-time visual querying of big data". CGF, 32(3):421–430, 2013.

[52] L. D. Lins, J. T. Klosowski, and C. E. Scheidegger, „Nanocubes for real-time exploration of spatiotemporal datasets". TVCG, 19(12), 2013.

[53] H. Wickham. Bin-Summarise-Smooth, "A framework for visualising large data", Technical report, 2013.

[54] J. Im, F. G. Villegas, and M. J. McGuffin, "Visreduce: fast and responsive incremental information visualization of large datasets", In BigData, 2013.

[55] Neil Biehn, "The missing v's in big data: viability and value", Wired, 2013, https://www.wired.com/insights/2013/05/the-missing-vs-in-big-data-viability-and-value/

[56] Bill Vorhies, How Many "V"s in big data – the characteristics that define big data", October 31, 2013 http://data-magnum.com/how-many-vs-in-big-data-the-characteristics-that-define-big-data/

[57] B. Shneiderman. "Extreme visualization: squeezing a billion records into a million pixels", SIGMOD, 2008.

[58] Judith Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman, Big Data For Dummies 1, "The validity, veracity, and volatility of big data", Wiley, 2015.

[59] Bill Vorhies, "How many "v"s in big data – the characteristics that define big data", October, 2013.

[60] Tim Harford, "Big data: are we making a big mistake?", Finiancial Times, March 28-2014, https://next.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0.

[61] P. Bedi, V. Jindal, and A. Gautam, "Beginning with big data simplified", International Conference on Data Mining and Intelligent Computing (ICDMIC), 1-7, 2014.

[62] Andra-Raluca, "Development of a Capability Maturity Model for Big Data Governance Evaluation in the Belgian Financial Sector", Master's Thesis - Master in Business Administration Graduation, Faculty of economics and business campus brussels, June 2015.

[63] V. Morabito, "Trends and challenges in digital business innovation", Springer International Publishing. Switzerland, 2014.