ISOLATING VALUES FROM BIG DATA WITH THE HELP OF FOUR V'S

V.S.Thiyagarajan[1], K.Venkatachalapathy[2]

[1]*Research Scholar, Department of Computer Science, Annamalai University, TN, India*
[2]*Professor, Department of Computer Science, Annamalai University, TN, India*

Abstract

Big Data refers to the massive amounts of data that collect over time that are difficult to analyze and handle using common database management tools. It includes business transactions, e-mail messages, photos, surveillance videos and activity logs. It also includes unstructured text posted on the Web, such as blogs and social media. Big Data has shown lot of potential in real world industry and research community. We support the power and Potential of it in solving real world problems. However, it is imperative to understand Big Data through the lens of 4 Vs. 4th V as 'Value' is desired output for industry challenges and issues. We provide a brief survey study of 4 Vs. of Big Data in order to understand Big Data and extract Value concept in general. Finally we conclude by showing our vision of improved healthcare, a product of Big Data Utilization, as a future work for researchers and students, while moving forward.

Keywords: *Big Data, Surveillance videos, blogs, social media, four Vs.*

----------***----------

I. INTRODUCION

Big Data is defined as large set of data that is very unstructured and disorganized. In light of our study, we define it as "lesser and lesser the understanding of data is, bigger and bigger it would become". Big Data has lot of potential and, it is true as long as size of data itself does not become the part of the problem [1].

According to many researchers and writers, big data is a form of data that exceeds the processing capabilities of traditional database infrastructure or engines. High volume, and high velocity and high variety of such data make it an unfit candidate to our currently employed and tested database architectures. For any industry to implement Big Data tools and technologies, it is imperative to understand what makes Big Data. Today, the research is directed, encouraged and supported to understand, analyze, clean, process and utilize it for specific purposes. Oracle, a big database enterprise systems solution, talks about three types of Big Data [2]. 1) Traditional enterprise data (CRM Systems, ERP data, web store and general ledger data). 2) Machine generated/ sensor data (Call Details Records (CDR), weblogs, manufacturing sensors, digital exhaust, trading systems data). 3) Social data (Facebook, twitter, blogs, emails, customer feedback, reviews). Social networking companies like FB and Twitter are utilizing the power of Big Data that the users generate every minute. In fact Data is in motion always.

However, Value is the most desirable output of Big Data processing. Therefore, we must understand all 4 V's of it and we must then extract value from it. Rests of sections are divided as follows: Section II gives motivation for this paper. Section III through Section IX discusses 7 V's. Section X talks about future work in light of what we discuss in this paper. Section XI Concludes the vision given in this paper. Figure 1 shows the architecture of Big Data.

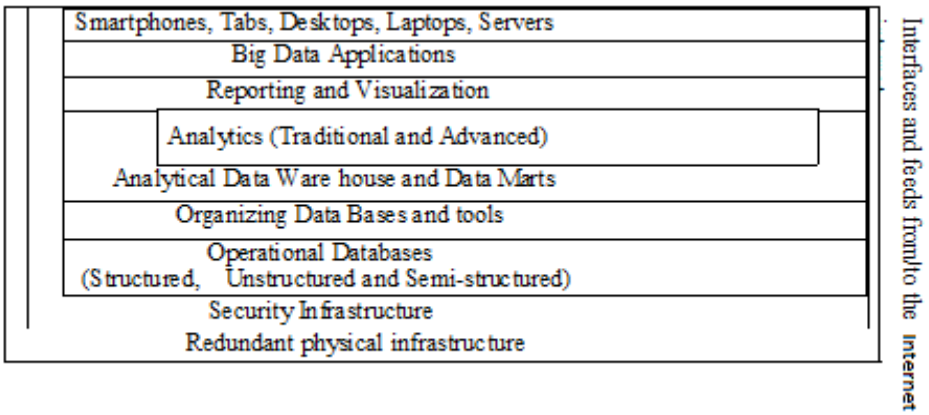

Fig 1: The Big Data Architecture

2. SYNOPTIC VIEW

Section III gives whole theme for this paper. Section IV through Section VII discusses 4 V's. Section VIII gives the future work in light of what we discuss in this paper. Section IX Concludes the vision given in this paper.

3. MOTIVATION

The motivation to write this paper and outlining relevant arguments comes from the fact that Big Data have become the part of our lives and Big Data hides in it the solutions to many problems in any industry. As a fact, Big Data provides raw ingredients to build tomorrow's great machines. We support the fact that Big Data will eventually take over the world of technology and internet. Big Data will play role to understand human as we all human are data agents. We all are generating data 24/7. Before we go out and look for big data, we must start within ourselves. We are the part of Big Data Ocean. We are generating data every second. When we read, we think, we write, we produce data. In a traditional concept, Data is everywhere.

In Public, Social networks, schools, hospitals, law enforcement agencies, entertainment and film making companies, governments, small and large businesses, weather data, climatic data, space exploration companies like NASA etc. With that said, we believe that there is a need to survey a data activity with some of these reputable companies including but not limited to Google, Amazon, eBay, LinkedIn, Facebook, Twitter, CNN, Weather Channel, etc. However that is beyond of the scope of this paper. In more technical view, Today's web is the main source of generating big data. We are spending more time on web than ever before. Still, lot of population in developing world has no access to web and they are not yet part of big data though. However, sooner or later, they will and Big Data issues will only get worse, if not taken/considered seriously today.

Today's web, including mobile apps is creating a huge trail of data that is by no means, understood but being thrown out there every second. We may draw some analogy of such growth with Moore's law. [3] Puts some bright lights on such discussion. According to study in [4], in year 2012, we have generated about 500 petabytes of Healthcare data. It is expected to grow to 25,000 petabyte by year 2020. According to Stephen Gold, VP of Marketing for IBM's Watson, we have created 90 % of all data in last 2 years and every day, we are adding to ocean of data by rate of 2.5 quintillion bytes of data [4]. His exact words are given as *"Big Data is the fuel. It is like oil. If you leave it in the ground, it doesn't have a lot of value. But when we find ways to ingest, curate, and analyze the data in new and different ways, such as in Watson, Big Data becomes very interesting."*

Google, with its search engine has become the greatest Data Company [5]. Google BigTable, Hadoop and MapReduce have revolutionized the industry and companies and provide great solution to Big Data real world problems [6][7]. Parallel and distributing computing model has given ability to perform complex operations on very large data sets. It deals with high volume, high velocity and high variety of data by bringing computation processing closer to data rather than brining data to computation as happened before big data era. Some of industry analyst has shown through studies that Big Data can possible contribute to various industries and market including but not limited to Healthcare, Job Market, stock market, retail, real estate, education, finance, environmental research, genomic, sustainability, politics and biological research.[8][9]. In essence, we have big data being produced anyways from all corners. It is a lot wiser to make intelligence and value out of it. Understanding and discussing all IV's in this paper will open doors towards finding true value of big data. We show our vision in the Figure 2 below.

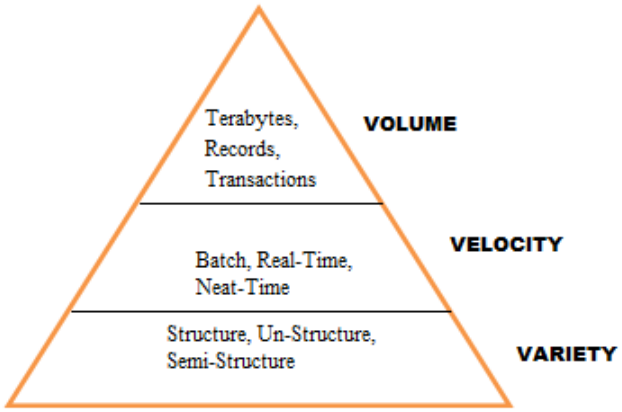

Fig.2. Three Vs. of Big Data

4. VOLUME 1st V

Big data implies enormous volumes of data [14]. It used to be employed created data. Now that data is generated by machines, networks and human interaction on systems like social media the volume of the data to be analyzed is massive. Volume of Big data refers to the size of data being created from all the sources including text, audio, video, social networking, research studies, medical data, space images, crime reports, weather forecasting and natural disasters, etc. According to [1], Input data to big data systems could be chatter from social networks, web server logs, traffic low sensors, satellite imagery, broadcast audio streams, banking transactions, MP3s of rock music, the content of web pages, scans of government documents, GPS trails, telemetry from automobiles, financial market data, the list goes on.

However such volume of data being, disorganized and unknown, cannot be handled or processed or queried with traditional ways, for example, SQL. In a simple example, you can no longer write a query like *"select something from some table where something equals something"*. Remember, unstructured data is no way near to be normalized in tables or data set that we are used to work with, in RDMS systems like Oracle and SQL Server. At the same time, we must

realize that we are dealing with peta byte of unstructured data here. SQL based approach simply does not work.

5. VELOCITY 2nd V

The velocity of big data, coupled with its variety, will cause a move toward real-time observations, allowing better decision making or quick action. As the market evolves, it is likely that most of these observations will be the result of custom applications designed to augment the ability to react to changes in the environment. Analysis frameworks and components will help to create, modify, share, and maintain these applications with greater ease and efficiency.[15]

Generally, it is the speed or velocity of data that makes it too much to work with. Imagine, the speed we are generating this kind of data using our Smartphone's and World Wide Web. It is absolutely nowhere near to be controlled as one might think. This high velocity is directly responsible for high volume that we spoke about in Section IV. With this high velocity of data coming in, businesses have to be prepared with technology and database engines to process them as they need them. Therefore, not only velocity to incoming data, that matters, but to stream the fast-moving data into big storage for later processing and analysis. Speaking of which, we talk about two important reasons for such data processing considerations. 1) To store input data since it is too fast at arrival. This requires some special analysis at time of data occurrence on the fly. 2) Application forces response to data as it arrives.

6. VAREITY 3rd V

As we know data appears in many shapes. Audio, video, text, images, you name it. This brings the real complexity to the mix. That is why we can't call it relational database any more. It is a great challenge to establish or build a system so such data mix can be integrated into it directly. Rarely does data present itself in a form perfectly ordered and ready for processing. A common theme in big data systems is that the source data is diverse, and doesn't fall into neat relational structures. It could be text from social networks, image data, a raw feed directly from a sensor source. None of these things come ready for integration into an application [16]. On the WWW, people use different software's, browsers and they send data differently to the cloud. Not to ignore, most data is coming directly from real human interface and errors are unavoidable in data. We find that Variety of data directly affects the integrity of data. In other words, more variety in data is: more errors it will contain.

7. VALUE 4th SPECIAL V

We call this V as Special for a reason. Unlike other V's of big data that we talked in earlier sections, this V is the desired outcome of big data processing. We are always interested to extract maximum value from any big data set we are given to work with. Here we must look for true value of data we are given to work with. In other words, data value must exceed its cost or ownership or management. One must pay attention to the investment of storage for data. Storage may be cost effective and relatively cheaper at time of purchase but such underinvestment may hurt highly valuable data, for example storing clinical trial data for new drug on cheap and unreliable storage may save money today but can put data on risk tomorrow [10]. Value of data greatly depends on governance mechanism as well. That is, how we write policies and structures that will eventually bring balance between reward and risk of the data [10]. Same time, these policies and structures, if not carefully written and implemented may restrict businesses to extract true value of data. In other words, it will make data undervalued. Another important point that is often ignored is that true value lies in the eyes of the customer of business data.

Another fact worth noting is that some of data a time of collection may not have same value to risk ratio but could develop as time goes on. [10] Shows some simulated results for Year 1 to Year 5 for % changes in hardware costs, non-hardware costs and total costs. With economy getting worse, IT budgets are being shrunk. Storage is always expensive. About 47 percent of IT budget to maintain IT infrastructure, 40 percent to information and transaction processing and about 13 percent to strategic IT investments [11]. Often data can migrate between various tiers. Higher the tier is, higher the value is. In other words, data at higher tiers will have lower risk of catastrophe. Therefore, some organization can accept high cost with storage associated at higher tiers as protection is better guaranteed at those levels and thus value to cost ratio is higher [12][13].

8. FUTURE WORK

After studying literature about big data research and current state of art and then summarizing ideas in this paper, we suggest the following future work in points below.

1. Designing special tools and techniques to extract value from Such a big stream data, which can be used to specific industry.
2. Developing very specific algorithms to utilize 3 Vs. of Big data in order to find 4th V.
3. Developing Healthcare solutions to utilize the big data in order to improve healthcare, disease control, disease early diagnosis and medicine production.
4. Algorithm needs to be written to develop evidence base medicine and personalized medicine by utilizing existing scientific evidence and test data.
5. Developing machine intelligence using cognitive science, AI and Big data to address availability of education to rural areas, to promote renewable energy and clean environments, to implement security measure to keep everybody safe, to improve industry and economy by creating more jobs and putting more people to work around the world, to control crimes and to solve poverty issues in developing world.

9. CONCLUSION

Big data has raw ingredients for tomorrow invincible machines. Data growth is promoting lot of technology innovation and creation. By understanding 4 Vs. of Big Data, we can utilize its power for specific research and real

world problems. To clarify matters, the four Vs. of *volume, velocity, variety and value* are commonly used to characterize different aspects of big data. They're a helpful lens through which to view and understand the nature of the data and the software platforms available to exploit them. Most probably you will contend with each of the Vs. to one degree or another. Therefore, research is highly needed in Big Data for many reasons discussed and information provided in this paper.

REFERENCES

[1]. Edd Dumbill (O'Reilly Media), "Volume, Velocity, Variety: What You Need to Know About Big Data"
[2]. Oracle: Big Data for the Enterprise – An Oracle White Paper June 2013
[3]. Big Data Now: Current Perspectives from O'Reilly Radar
[4]. Big Data in Healthcare - Hype and Hope, Bonnie Feldman, Ellen M. Martin, Tobi Skotnes October 2012.
[5]. Big Data in Big Companies, May 2013, Thomas H. Davenport, Jill Dyche. International institute for Analytics.
[6]. The Apache Software Foundation. http://hadoop.apache.org/common/credits.html
[7]. Ghemawat D.J .MapReduce: simplified data processing on large clusters. In: Proc of OSDI, 2004
[8]. IDC, Digital Universe in 2020
[9]. Nasscom-CRISIL GR&A Analysis, Reuters
[10]. Judith Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman – Big Data for Dummies. ISBN: 978-1-118-50422-2
[11]. Paul P. Tallon, Lyala Universtiy Maryland – Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost, IEEE Computer Society 2013.
[12]. P. Weill, S.L. Woerner, and H.A. Rubin, "Managing the IT Portfolio (Update Circa 2008): It's All about What's New," MIT Center for Information Systems Research (CISR), vol. 8, no. 2B, 2008, pp. 1-4.
[13]. P.P. Tallon, R.V. Ramirez, and J.E. Short, "The Information Artifact in IT Governance: Towards a Theory of Information Governance," to appear in J. MIS, Loyola Univ. Maryland, 2013
[14]. Inside Big Data Group http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/
[15]. A Wiley Brand, "Big Data for Dummies", A Wiley Brand publication by John Wiley & Sons, Inc. 2013 edition.
[16]. O'Reilly, "Big Data Now" by O'Reilly Media, Inc. 2012 Edition.

BIOGRAPHIES

V.S.THIYAGARAJAN obtained his Bachelor's and Master's degree in Computer Science Engineering from Department of CSE, Annamalai University. Then he obtained his Master's degree in Business Administration from the same university. He has also obtained Post Graduate Diploma in Data Mining from Directorate of Distance Education, Annamalai University. Currently he is doing his research in Big Data Analytics.

Dr. K.VENKATACHALAPATHY received his Master's degree in Computer Applications from Pondicherry University in 1990. He completed his Ph.D. in Computer Science & Engineering from Annamalai University, Tamilnadu. He is currently working as Professor in the Department of Computer Science & Engineering, Annamalai University