Seven V's of Big Data

Understanding Big Data to extract Value

M. Ali-ud-din Khan, Muhammad Fahim Uddin, Navarun Gupta

***Abstract*—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 7 V's. 7th V as 'Value' is desired output for industry challenges and issues. We provide a brief survey study of 7 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, Unstructured Data**

I. INTRODUCTION

Most commonly, 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. In the beginning of big data era, it was believed that the main contributors to big data were scientist and physicists, military experiments and simulations, NASA and super computers. More recently, busy traffic of airlines, advance healthcare system and insurances, banking and stock exchanges and electronic money mechanisms and systems contribute. However, Value is the most desirable output of Big Data processing. Therefore, we must understand all 7 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.

II. 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.

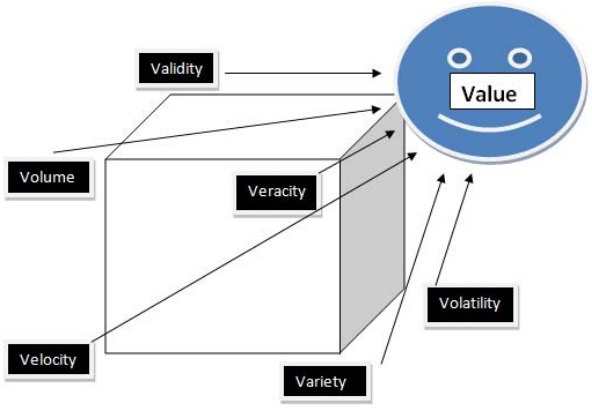
Today's web, including mobile apps is creating a huge trail of data that is by no means, understood but being thrown out there every second. We may draw some analogy of such growth with Moore's law. [3] Puts some bright lights on such discussion.

978-1-4799-5233-5/14/$31.00 ©2014 IEEE

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 veracity 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 V's in paper will open doors towards finding true value of big data.
We show our vision in the Figure 1 below.

FIGURE 1 - 6 V'S OF BIG DATA EVOLVING INTO VALUE OF DATA IN IT.



III. 1ST V – VOLUME

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, etc. According to [1], Input data to big data systems could be chatter from social networks, web server logs, traffic flow 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.

IV. 2ND V – VELOCITY

This is another important term to talk about when it comes to big and complex data. 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 III. 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. There is an interesting point in commercial form IBM as "*you wouldn't cross the road if all you had was a five-minute old snapshot of traffic location. There are times when you simply won't be able to wait for a report to run or a Hadoop job to complete*." That means that important thing is the speed of feedback loop that takes data from input through to the decision. 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.

V. 3RD V – VAREITY

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. 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.

VI. 4TH V – VERACITY

By Veracity, we mean the truthfulness of data. In other words, how certain we are about this data? Or how much data is the kind, it claims to be of that kind? We are most likely talking about meaningfulness of results from data for given problem space, we are working on or exploring.

Before this V was added to big data universe, it was assumed in the scientific and research community that incoming data is clean and precise. This assumption was actually followed very much in traditional data warehouses.

Now think about it. We are dealing with unstructured and big data here, which might be coming from Facebook posts, tweets, LinkedIn posts, etc. Do we trust whatever we see out there? Though we enjoy such data posts while reading and contributing, but we may not realize that we cannot rely on it for our sales and business or critical decision.

In my view this V is of highest concern to the processing of big data and related analysis and results outcome. Remember, we do normalization to our relational and traditional database to sustain the integrity of data. The data, which we can trust and it has no duplicates. Therefore, it is of vital importance to consider the cleansing of big data with some great tools and algorithms. Also, we need to come up of our definition of trust for such data depending upon the source and possible consumption of this data, however.

VII. 5TH V – VALIDITY

Validity of data may sound similar to veracity of data. However, they are not the same concept but similar. By validity, we mean the correctness and accuracy of data with regard to the intended usage. In other words, data may not have any veracity issues but may not be valid if not properly understood. Critically speaking, same set of data may be valid for one application or usage and then invalid for another application or usage. Even though, we are dealing with data where relationship may not be defined easily or in initial stages, but it is very important to verify relationship to some extent between elements of data, which we are dealing with to validate it against intended consumption, as possible.

As an example given in [3], Can a physician simply take a data from clinical trial that is related to patient's disease symptoms without validating them? The answer is No.

Another example from [3], we can verify or validate the storm potential in some areas predicated by Weather Satellites along with Tweets to see how much impact is going to be on individuals.

VIII. 6TH V – VOLATILITY

Speaking of volatility of big data, we can easily recall the retention policy of structured data that we implement every day in our businesses. Once retention period expires, we can easily destroy it. As an example: an online ecommerce company may not want to keep a 1 year customer purchase history. Because after one year and default warranty on their product expires so there is no possibility of such data restore ever. Big data is no exception to this rule and policy in real world data storage. Such issue is very much magnified in big data world and not as easy as we have dealt with it in traditional data world. Big data retention period may exceed and storage and security may become expensive to implement. Actually, Volatility becomes significant due to Volume, Variety and Velocity of data.

IX. 7TH SPECIAL V – VALUE

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].

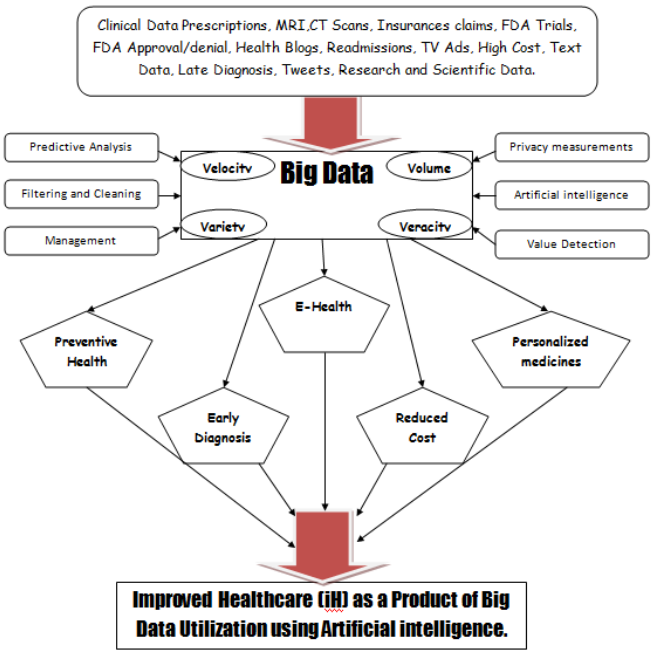
X. FUTURE WORK

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 use to specific industry.

2. Developing very specific algorithms to utilize 6 Vs of Big data in order to find 7th 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.
6. Finally one of the prime promises of Big Data is Healthcare solutions. In Figure below, we show our vision of Big Data Producing improved healthcare using artificial intelligence.

FIG 2 – IMPROVED HEALTHCARE EXAMPLE UTILIZING BIG DATA.



XI. CONCLUSION

Big data has raw ingredients for tomorrow invincible machines. In Healthcare alone, By understanding, processing and utilizing the knowledge and information hidden in Big Data with respect to health issues, disease trends in certain population, we can find hidden fountain of youth, with which, we can live longer and healthier. Data growth is promoting lot of technology innovation and creation. By understanding 7 Vs of Big Data, we can utilize its power for specific research and real world problems. As a future work, one of outstanding research direction we recommend is Healthcare. Without Big Data in Healthcare, we are at mercy of Doctor's experience that may not work accurately in all cases and literature that may not be quickly available. However, With Big Data and Intelligent Algorithms, we can eventually grow the field of predictive and personalized medicine that will make efficient use of case studies, treatment histories, and disease and prescription data and finally will improve healthcare and cost to the point where we all expect it be.

Even though there is a fear about unintended harm, Big Data can bring, but it is believed and forecasted that the benefits, at the end of the day, will outweigh such harms and therefore, research is highly needed in this direction 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

M. Ali-ud-din khan is a recent graduate from University of Peshawar, Pakistan. He has over 15 years of experience in Pharmaceutical companies holding various technical and managerial positions. He also holds MBA degree.

He is currently pursuing a project HRP-ADH (Health Recovery Program-Aktion Deutschiand Hilft) German, working in Pakistan with a company named as CAMP/HELP (Community Appraisal & Motivation Programme).

His interests are Big Data and Healthcare. At present, Mr. Khan is developing his real world skills through related projects in healthcare and reading literature about it. He is looking forward to attending conferences and publishing papers to explore industry and research community. He looks forward to pursuing his PhD in future.

Muhammad Fahim Uddin is a PhD Student at University of Bridgeport in Connecticut, USA. He holds a bachelors (UET Peshawar) and Masters (University of Bridgeport) in Electrical Engineering.

Mr. Uddin is currently researching Big Data in Healthcare. His focus is to understand Big Data, explore its potentials to solve real world industry problems, particular in area of Healthcare and Preventive/Personalized medicine. He envisions an Intelligent Healthcare (iH) as a product of Utilizing Big Data using Artificial intelligence techniques and algorithms. He has been working in IT as a Data Architect/DBA with local Government, since 2007. His research interests are mainly focused on Big Data intelligence, Big Data predictive and preventive abilities, Artificial intelligence and Machine learning. His goals of research are to extract Value from Big Data in order to improve Healthcare and related real world problems, through his PhD research work.

Navarun Gupta is an Associate Professor and Chair of Electrical Engineering at the University of Bridgeport in Connecticut. He holds a Ph.D. in Electrical Engineering from Florida International University, a master's in Physics from Georgia State University and a master's in Electrical Engineering from Mercer University.

Dr. Gupta's interests include audio and bio signal processing. Besides teaching, he supervises several master's theses and is advising one Ph.D. student. He is also an active member of the biomedical engineering program at the University of Bridgeport. Gupta also likes to work with the local schools in the area of Bridgeport to encourage students to take up engineering as a career. He and his graduate students have been working with middle school students in Bridgeport to improve computer literacy. They are also involved with the Project Lead The Way program at Stratford High School in Connecticut.

As a past Chair of ASEE (American Society for Engineering Education), Dr. Gupta has been very active in promoting engineering education in high schools. He is the organizational chair of Zone 1 of ASEE Conference that will be held at University of Bridgeport on April 3-5, 2014