*International Journal of Latest Trends in Engineering and Technology*
*Vol.(8)Issue(2), pp.063-066*
*DOI: http://dx.doi.org/10.21172/1.82.009*
*e-ISSN:2278-621X*

# THE NEW SIXTH 'V' OF BIG DATA AND WEB INTELLIGENCE

Dr. Rajiv Chopra[1]

**Abstract- Data may be acquired by people, from machines or from the web. Present day is an era of data lakes. Organizations must be able to store, manage and manipulate vast amount of data at right speed and at right time to gain right insights. This paper combines the Artificial Intelligence (AI) and Big Data together that has added a new feather in the field of research i.e. the Web Intelligence. AI techniques have been applied on Big Data that is growing exponentially every second. Further-more, it is Big Data Analytics and Web Analytics that has made it all possible.**
**Keywords: Artificial Intelligence, Big Data Analytics, Big Data, Web Analytics, Web Intelligence**

## I. INTRODUCTION

Web has resulted in deluge (flood) of data. Today data is the fuel of growth and innovation. Data may be structured (like RDBMS) or unstructured (like twitter data), real or non-real, data at rest (static data) or data in motion (dynamic data). But data is very informative. The challenge is to save it and then filter out only relevant and mandatory data. This is so because of its sheer volume. Web consists of millions of servers with data store of sizes in zetabytes (1 zetabyte = 1000 patabytes; 1 petabyte = 1000GB; 1 exabyte = 1000 patabytes). Big data refers to the voluminous data which resembles to tha of a gold mine. A gold mine has lesser gold but other things are more. Similarly, relevant data is less in Big Data but other types of data are more. Thus, analytics of Web and Big Data is a must. Big Data Analytics (BDA) refers to the process of examining data, typically of a variety of sources, types, volumes and/or complexities-to uncover hidden patterns, unknown correlations and other useful information [1]. Web analytics is the measurement, collection, analysis and reporting of web data for the purpose of understanding and optimizing web usage [1-2]. It is not just a tool for measuring web traffic but can also be used as a tool for business and market research and to assess and improve the effectiveness of a web site.

According to the Kryder's Law-"Storage capacity per dollar is growing much faster rather than the computing power per dollar that appears to double every 18 months (Moore's Law)." So, both the computing power as well as the raw data are available now. The collaborative use of AI techniques at web scale on several millions of servers has resulted in giving a basic form of intelligence called as 'human intelligence'. It is the application of web-scale computing power on voluminous big data (due to Internet) has given birth to 'Web Intelligence'. It can be shown in form of an equation too:-

A.I. + BIG DATA = WEB INTELLIGENCE (1)

## II. LITERATURE REVIEW

Based on several survey results some of the studies are as follows:-

- Web may have more than a trillion of web pages. Out of these, 50 billion are already indexed by Google (via Search Engines).
- Huge web content may cross 100 million domains (locations where we point browsers).
- Facebook and Twitter have over 900 million users each. They generate more than 300 million posts / day. Then over 10,000 credit card payments made per second, over 6 billion mobile phones, images and videos on YouTube and other sites has resulted in sheer volume of data.
- Vannevar Bush [1] et al in 1945 emphasized on the fact that effort should be more on emulating and augmenting human memory. They even created a MEMEX device modeled on human memory.

[1] *Department of Computer Science and Engineering Guru Tegh Bahadur Institute of Technology Afft .Guru Gobind Singh Indraprastha University, New Delhi, Delhi, India*

- Recently, Google introduced 'Google Glass'. It just needs users 'look-through' and it would retrieve information about it. Even images may be used for look-up.
- Technology has today progressed a lot wherein we just need to 'look up' the global collective memory.
- Google handles more than 4 billion search queries per day.
- Human brains are quite poor at indexing. So, we search the web. Google's million servers regularly crawl and index over 50 billion web pages (estimate).
- Web pages are stored based on some model of memory so that the PageRank algorithm can be applied. Researchers suggest a 'semantic model' for the same. In this model, pairs of words are associated with each other in some way as being synonyms of each other or one a generalization of the other. So, we can put it in an equation form:-

WWW = WEB PAGES + HYPERLINKS (2)

SEMANTIC MODEL = WORDS + ASSOCIATIONS (3)

- As per the survey report of Brown University, it used a semantic model of about 5000 common words i.e. a network of word-associations pairs. They executed the PageRank algorithm using the network of web pages and hyperlinks. This produces a ranking of all the words by importance.
- Experiments in 2007 reports that roughly 2.5% of random sample of web pages were forms that should be considered as part of deep web. Google approach is to automatically try out many possible inputs and input combinations for a deep web page and to find out that gives better results. These results are then stored internally in Google index. This makes them part of the surface web.
- Kosmix (2010) classified popular web-based services considering both automated and human assisted processes. Searching deep-web is one of the current areas of research today.
- In another report, the FBI investigator, Robert Fuller searched a commercial database (named as CheckPoint) to search for some terrorist of 9/11 attacks. This database stored personal information on US residents like their mobile numbers and addresses. Later a journalist, Bob Wood ward quoted that if FBI would have done a simple credit card check then they would have found that two men bought 8 tickets for morning flights. Now this is web knowledge and it would have stopped the attacks.
- Data may be structured or unstructured. In its crude form, data has lesser value. But if data is corrected for errors, aggregated, normalized, calculated or categorized then its value grows dramatically.
- Till 2012, 2.5 exabytes (approx.) of data is created each day. This volume comes from both new data variables and the amount of data records in those variables.
- The outcome is now to possess mountains of data that can provide the building blocks to information generation through analytics.
- As per the latest IBM report-
  - 2.5 petabytes is memory capacity of human brain.
  - 13 petabytes is the amount that could be downloaded from internet in 2 mins., if every American (300M) got on a computer at the same time.
  - 4.75 exabytes-total genome sequences of all people on Earth about 7 billion in 2015.
  - 422 exabytes- total digital data created in 2008.
  - 1 zetabyte- world's digital storage capacity in 2013.
- In another report by IDC-
  - Digital Universe is doubling in size every two years.
  - By 2020, the digital Universe will reach 44 zettabytes.
- Between 2013 and 2020, the division of the digital Uinverse between mature and emerging markets will switch from 60% accounted for by mature markets to 60% of the data in digital Universe coming from emerging markets.
- As of 2012, everyday 2.5 exabytes (i.e. $2.5 * 10^{18}$) of data are created. IDC and EMC [5] project that data will grow to 40 zettabytes by 2020, resulting in a 50-fold growth from the beginning of 2010.
- The results of a global study by CA Technologies has revealed that Enterprise Big Data strategies are delivering key benefits to organizations. 44% of the Indian respondents have already implemented a Big Data project while 36% plan to implement one in the coming year. Also 12% of the Indian respondents have stated that they have implemented four Big Data projects to date.
- The study, titled The State of Big Data Infrastructure lists five major obstacles for successful implementation of Big Data, insufficient existing infrastructure (33%), organizational complexity (25%), security concerns (28%), lack of budget/resources (25%) and lack of visibility into information and processes (21%).
- Big Data strategy will improve customer experience (67%), need to enter new markets (52%) and need to get new customers (49%).

- As per the survey done by EY on Big Data and enterprise mobility in India with 110 IT leaders across industry sectors to understand the relevance of Big Data and enterprise mobility in the Indian context:-
  - More than 60% of organizations define Big Data as large volumes of unstructured data. Less than 20% of organizations consider Big Data as "hype" or the latest technology buzzword.
  - About 60% acknowledge that unstructured information is growing out of control and is driving Big Data explosion, with unstructured data expected to grow at a faster rate (30-40%) than structured data (30%) over the next 3-5 years.
  - Half of the large organizations plan to use data from social networks for sentimental analysis and customer tracking.
  - About 80% of organizations are in early stages of Big Data initiatives and 60% are still in their infancy stage.
  - 75% of organizations are confident of driving new revenue streams using Big Data. However, only 35% plan to invest in building Big Data capabilities related to analytics, security software and real-time applications.
- There is a huge potential for Big Data technologies in India. This is a high growth economy and organizations depend on insights from raw data, to plan their future course. Big Data Analytics helped in parts, responsible for NDA to win General Elections in 2014.
- Several global Big Data players have sprouted in India over the past two years. Sears, for instance, established a wholly-owned subsidiary MetaScale to service healthcare and entertainment customers with revenues between US$1 million and US$ 10 million. A @WalmartLabs facility also opened in Bangalore in April to develop social media and analytics and Big Data infrastructure. In July, Yahoo set up a grid computing lab at IIT Madras campus.
- eBay[4] uses two data warehouses at 7.5 petabytes and 40PB as well as a 40PB Hadoop cluster for searching.
- Amazon.com handles millions of back-end operations every day as well as the queries from more than half a million third-party sellers. Amazon , by 2005, had the world's three largest Linux databases with capacities of 7.8 TB, 18.5 TB and 24.7 TB.
- In 2012, the Big Data Research and Development Initiative, was to find how Big Data could be used to address important problems faced by the government. The Utah Data Center is a data center currently has the facility to handle an exact amount of storage space is unknown but more recent sources claim it will be of the order of a few exabytes.
- Facebook handles 50 billion photos from its database. Google was handling roughly 100 billion searches per month, till August 2012.

## III. PROPOSED WORK

Web Data, Web Information, Web Knowledge and Web Intelligence are related as depicted in figure-1.

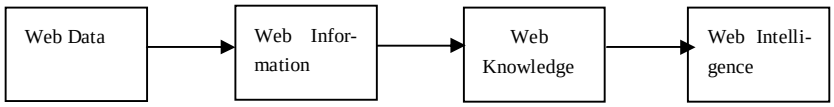


Figure 1. Relation between Web Data, Information, Knowledge and Intelligence

Raw web data when processed becomes more meaningful information which when processed further becomes web knowledge. In web knowledge, lays the power!! Web knowledge when further processed, it becomes Web Intelligence.

The complexity of this data has resulted in another factor- *Vulnerability of Big Data*.
The researchers have already explored on five (5) V's of Big Data i.e. Volume, Velocity, Veracity, Variety and Value. The complexity of Big Data has given birth to another V i.e. Vulnerability of Big Data. By vulnerability, it means that the web data can be easily stolen if its security level is not high.

As 'machines' work on 'thinking', similarly web-based engines are focal points today. Web knowledge, so obtained can result in more of web intelligence. So, web –intelligence programs that have their own ability to understand us and this world are needed. A web intelligence program looks at the data stored in or moving on Internet. Web intelligent systems learn about our preferences and behavior.

Newly developed stf-driving cars are possible as web provides us both information and a communication platform. Web intelligence systems will evolve synergistically with our social intelligence using web itself.

No inference is possible in absence of right input data. Such logical inferences may be automatically conducted by machines while the knowledge may be learned from experience.

On millions of servers, several copies of web index are stored. This increases processing speeds. The information on web that is indexed by the search engines (like Google) is called as the 'surface web'. On the other hand, 'deep web' consists of data hidden behind web-based services. The point is that the volume of data within deep web is huge, exponentially large and of infinite size. Also it is to be pointed out here that this deep web is huge and even bigger than an indexed web of 50 billion pages.

## IV. CONCLUSIONS

We are trying to look at the world through new lenses but at the same time the lenses themselves are looking back to us and this is what we will proceed to explore next. Data are the building blocks to information and information is vital for knowledge generation for decision makers across industry to make better decisions. Healthier decisions means greater operational efficiencies, cost reductions and reduced risks. A new 'V' in this dimension must be taken care of as even big data on web servers may get hacked easily.

## REFERENCES

[1] G. Shroff, "The Intelligent Web: Search, Smart Algorithms and Big Data", *Oxford University Press, 2013.*

[2] Rajiv Chopra, "Testing Web Applicatons: The State of the Art and Future Trends",*Cambridge International Science Publishing, UK*, 2016.

[3] Rajiv Chopra, "Web Engineering", Prentice Hall of India (PHI), 2016.

[4] J. S. Hurwitz et al.,"Big Data for Dummies", John Wiley & Sons, Inc., 2013.

[5] Rajiv Chopra, "Cloud Computing", New Age International Publishers, 2016.