International Journal of Engineering Technology, Management and Applied Sciences

IJETMAS

www.ijetmas.com May 2016, Volume 4, Issue 5, ISSN 2349-4476

A New Approach to Big Data Challenges and Opportunities

Madhavi Tota[1]
Information Technology Dept.
Rajiv Gandhi College of Engineering
Research and Technology. Chandrapur

Dr. Chandrashekar A. Dhote[2]
Professor, Information Technology Dept.
VYWS. Prof. Ram Meghe Institute of
Technology & Research, Badnera

Abstract:

Innovations in technology and greater affordability of digital devices have presided over Today's Age of Big Data, an umbrella term for the explosion in the quantity and diversity of high frequency digital data. These data hold the potential—as yet largely untapped to allow decision makers to track development progress, improve social protection, and understand where existing policies and programmes require adjustment. Turning Big Data—call logs, mobile-banking transactions, online user-generated content such as blog posts and Tweets, online searches, satellite images, etc.—into actionable information requires using computational techniques to unveil trends and patterns within and between these extremely large socioeconomic datasets. Heterogeneity, scale, timeliness, complexity, and privacy problems with Big Data impede progress at all phases of the pipeline that can create value from data. Much data today is not natively in structured format; for example, tweets and blogs are weakly structured pieces of text, while images and video are structured for storage and display, but not for semantic content and search: transforming such content into a structured format for later analysis is a major challenge. In this paper we are trying to bring together diverse perspectives, coming from different geographical locations with different core research expertise and different affiliations and work experiences. The aim of this paper is to evoke discussion rather than to provide a comprehensive survey of big data research and to suggest ways to address at least a few aspects of each heterogonous challenges.

Keywords: Privacy, Analytics, big data, data governance, data integration.

I. Introduction

"Big data" has become a popular theme in digital era and an important factor for production, development and growth of nation. But what exactly the term big data and on what factors is it depends. "Big Data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze [3].The datafication of our world includes many factors such as[12]:

- Activity Data: Music players,eReaders and smart phones collect data on how we use them, web browser collect information on what we search for credit card companies collect data on where we shop and shops collect data on what we buy.
- Conversation Data: Our conversations are being captured from emails to all the conversations we have on social medie sites like Facebook or Twitter as well as our phone conversations are now digitally recorded.
- Photos and video image Data: All the pictures and videos we take on our smart phones and digital cameras, we upload and share millions of them on social media sites every second.
- Sensor Data: we are surrounded by sensors that collect and share data devices like our smart phones use sensors to track our location, the speed and direction at which we are travelling , read our fingerprints detect how light it is outside, etc.

Big Data characteristics and their issues:

1. Volume: it refers to the vast amount of data generated in every second especially machine-generated data, is exploding, how fast that data is growing every year, with new sources of data that are emerging. Today we create the same amount of data in a single minute that was created from the beginning of time the year 2000. The challenge is how to deal with the size of Big Data and the value of different data records will decrease in proportion to age, type, richness and quantity.
2. Velocity: it defines the speed at which new data is generated and the speed at which data moves around. "As businesses get more value out of analytics, it creates a success problem— they want the data available faster, or in other words, want real-time analytics. And they want more people to have access to it, or

in other words, high user volumes." the key challenges is how to react to the flood of information in the time required by the application [3].

3. Variety: it is all about the different types of data we can now use. Today we don't have to rely on nicely structured data, we can now collect and analyse text, images, videos, voice, location data, and much more. How can we cope with uncertainty, imprecision, missing values, misstatements or untruths? It is probably the biggest obstacle to effectively using large volume of data.

4. Veracity: it refers to the messiness or trustworthiness of the data. Today quality and accuracy of data are less controllable (hash tags, abbreviations, typos and colloquial speech) but technology now allows us to deal with it. How to find high-quality data from the vast collections of data that are out there on the Web.

5. Value: the final V refers to the need to turn our data into value. Today big data is used to better understand and target customers, understand and optimize business processes, and improves health care, security and law enforcement. But the possible applications of big data are endless.

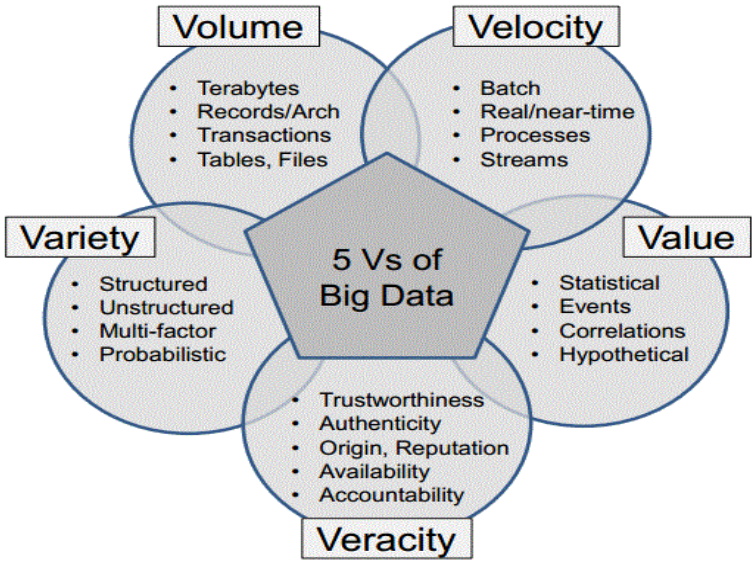

Fig 1: 5 V's of Big Data

The paper is organized as follows. Section I introduction to big data and its 5 V characteristic issues. Next, section II focus on day to day big data challenges in various areas. In Section III were discussed about different approaches for big data challenges. Section IV is all about different opportunities available for big data issues. The paper concludes with the summary and suggestions for further research.

II. Challenges in big data

The challenges in big data are usually the real implementation hurdles which require immediate attention. Any implementation without handling these challenges may lead to the failure of the technology implementation and some unpleasant results [15].

1. Privacy and Security

It is the most important challenges with big data which is sensitive and includes conceptual technical as well as legal significance.

i. The personal information of a person when combined with external large data sets, leads to the inference of new facts about that person and the person might not want the data owner to know or nay person to know about them.

ii. Information regarding the people is collected and used in order to add value to the business of the organization. This is done by creating insights in their lives which they are unaware of.

2. Data Access and Sharing of Information

If the data in the companies information system is to be used to make accurate decisions in time it becomes necessary that it should be available in accurate, complete decisions in timely manner. This makes the data management and governance process bit complex adding the necessity to make data open and make it available and format to better decision making.

3. Analytical Challenges

The typical analysis to be done on this huge amount of data which can be unstructured, semi structured or structured requires a large number of advance skills. Moreover the type of analysis which is needed to be done on the data depends highly on the result to be obtained i.e. decision making. This can be done by using two techniques: either incorporate massive data volumes in analysis or determine upfront which big data is relevant.

4. Human Resource and Manpower

Since Big data is at its youth and an emerging technology so it needs to attract organization and youth with diverse new skill sets. These skills not to be limited to technical ones but also should extend to research, analytical, interpretive and creative ones. These skills need to be developed in individual hence requires training programs to be held by the organization.

5. Technical Challenges

a. Fault Tolerance: with the incoming of new technologies like Cloud computing and Big data it is always intended that whenever the failure occurs the damage done should be within acceptable threshold rather than beginning the whole task from the scratch. Fault tolerant computing is extremely hard, involving intricate algorithms. It is simple not possible to devise absolutely foolproof. Thus the main task is to reduce this probability of failure to an acceptable level.

b. Scalability and complexity: Aggregating multiple disparate workloads with varying performance requires a high level of sharing of resources which is expensive and also brings with it various challenges. For data analysis, instance leveraging the influencers in a social network to create better user experience are hard problems to solve at scale[3]. All of these problems combined create a perfect storm of challenges and opportunities to create faster, cheaper and better solutions for Big Data analytics than traditional approaches can solve.

c. Heterogeneity and Incompleteness: the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step in (or prior to) Data analysis. However, computer systems work most efficiently if they can store multiple items that are all identical in size and structure. Efficient representation, access, and analysis of semi---structured data require further work.

III. Different approaches for big data challenges

a. Natural Language Processing (NLP) and Bioinformatics for Big data:

To handle Big Data in bioinformatics, biomedical informatics and biology (and Big Data in general), we would need some automated method as it is not possible for human to manually try to process, understand and derive new inferences from such large amount of data. Big Data consists of unstructured (free text data) and structured data (e.g. data in a database). Unstructured data dominates the data world. It is estimated that over 80% data in computers and Internet are unstructured [6]. Semantics is also very key to improve the usage of structured data – in finding relations, extracting new information and connecting / using structured data with unstructured data [8]. Thus, Natural Language Processing (NLP) and associated semantics become very useful in addressing Big data problems in bioinformatics and biology. In fact, use of NLP in biology has been increasing rapidly. A very good description of how NLP is used for Information Management in biology and bioinformatics is provided. From information management standpoint, NLP has 3 aspects: information retrieval, information extraction and semantics. Information retrieval refers to the recovery of documents from databases related to user's query. Search from the Internet and databases can be grouped under Information Retrieval. The goal is to find the most related information to the query.

b. Semantics-Empowered Approaches to Big Data: The key to handling volume is to change

The level of abstraction for data processing to information that is meaningful to human activity, actions, and decision making. We have called this *semantic perception,* which involves semantic integration of large amounts of heterogeneous data and application of perceptual inference using background knowledge to abstract data and derive actionable information. Our work involving Semantic Sensor Web (SSW) and IntellegO , which is a model of machine perception, integrates both deductive and abductive reasoning into a unified semantic framework that not only enables combining and abstracting multimodal data but also enables seeking relevant information that can reduce ambiguity and minimize incompleteness, a necessary precursor to decision making and taking action. Specifically, our approach uses background knowledge, expressed via cause-effect relationships, to convert low level data into high-level actionable abstractions, using cyclical perceptual reasoning involving predictions, discrimination, and explanation. Semantic approach can be used to solve 5 V characteristic problems of big data i.e. *Volume* by enabling abstraction to achieve semantic scalability (for decision making), *variety* by overcoming syntactic and semantic heterogeneity to achieve semantic integration and interoperability, *velocity* by enabling ranking to achieve Semantic filtering and focus, *veracity* by cross checking multimodal data with semantic constraints, and *value* by enriching semantic models to make them more expressive and comprehensive.

IV. Opportunities in real world applications

Fig 2: Big data application areas

Big Data can generate financial value across sectors. They identified the following key sectors:

a. Health care: (this is a very sensitive area, since patient records and, in general, information related to health are very critical) When combined with outcomes, high-quality data provided by patients can become a valuable source of information for researchers and others looking to reduce costs, boost outcomes and improve treatment. Several challenges exist with self-reported data[19][13]. Such as:

- Privacy concerns: People are generally reluctant to divulge information about themselves because of privacy and other concerns [11]. Creative ways need to be found to encourage and incent them to do so without adversely impacting data quality.
- Consistency: Standards need to be defined and implemented to promote consistency in self-reported data across the healthcare system to eliminate local discrepancies and increase the usefulness of data.
- Facility: Mechanisms based on e-health and m-health — such as mobility and social networking — need to be creatively employed to ease members' ability to self-report.

b. Financial Trading: high frequency trading is an area where big data finds a lot of use today. Here, big data algorithms are used to make trading decisions. The majority of equity trading now takes place via data algorithms that increasingly take into account signals from social media networks and news websites to make buy and sell decisions in split seconds.

c. Global personal location data [15]: (this is very relevant given the rise of mobile devices) big data analytics help machines and devices become smarter and more autonomous. Big data tools are used to operate Google's self- driving car. Big data tools are also used to optimize energy grids using data from smart meters.

d. Manufacturing: Manufacturing companies have long lists of production and shipping metrics. But what is needed now more than ever is the ability to obtain actionable information from these growing sources of data [7]. The following is a list of top areas Big Data technologies can impact in a manufacturing organization:

- Improved forecasting of products and production
- Enhanced service and faster support of customers
- Real-time decisions and alerts based on manufacturing data
- Integrated manufacturing and business performance information for improved decision making
- Rationalization of performance data across multiple plants

Analysis of supplier performance and better interaction and negotiations with suppliers

e. Social personal/professional data: Social media analytics (SMA) involves the collection of data from social media sites/applications (such as, wiki, Facebook, Twitter, GooglePlus, blogs etc) and evaluating such data to gain insights/knowledge [20].

- Almost 40% of social media users have bought an item after "favoriting" or sharing it on a social media site.
- 71% of social media users are more likely to purchase based on referrals.
- 74% of consumers depend on social media networks when making decisions on what to purchase.
- Facebook influences 30.8% of social media users purchasing habits, while LinkedIn and YouTube
- influence 27% of social media site users respectively

V. Conclusion

Big Data will not simply replace the approaches, tools and systems that underpin development work. What it does say, however, is that Big Data constitutes an historic opportunity to advance our common ability to support and protect human communities by understanding the information they increasingly produce in digital forms. Through Better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error---handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These Technical challenges are common across a large variety of application domains, and therefore not cost---effective to address in the context of one domain alone. Furthermore, these challenges will require Transformative solutions, and will not be addressed naturally by the next generation of industrial Products.

VI. Future scope

Fault Tolerance can be reduced by two ways first is to divide the whole computation being done into tasks and assign these tasks to different nodes for computation. Second, one node is assigned the work of observing that these nodes are working properly. If something happens that particular task is restarted and for recursive computations Checkpoints are used to keep state of the system at certain intervals of time. In any case, the computation can restart from last checkpoint maintain. For converting large heterogeneous volume of data into productive and value information, different approaches were purposed they are Natural Language Processing with Bioinformatics and semantic analysis.

VII. References

[1] Hanlon A. Big data and the 5Vs. 2014. Available at http:// hiveintelligence.com/big-data-5vs/ (accessed 24 January 2015)
[2] Salas-Vega et al.: Big Data and Health Care: Challenges and Opportunities for Coordinated Policy Development in the EU :Health Systems & Reform, Vol. 1 (2015), No. 4
[3] *Big Data: Challenges and Opportunitiespresented by Roberto V. Zicari in Big Data Computing*
[4] A visualization representing global Google search volume by language. Developed by the Google Data Arts Team. http://data-arts.appspot.com/globe-search : Big Data for Development: challenges and opportunities in May 2012.
[5] Challenges and Opportunities with Big Data *A community white paper developed by leading researchers across the United States*
[6] Big Data Opportunities and Challenges:Discussions from Data Analytics Perspectives Zhi-Hua Zhou, Nitesh V. Chawla, Yaochu Jin, and Graham J. Williams : IEEE COMPUTATIONAL INTELLIGENCE MAGAZINE, VOL. XX, NO. X, XXXX 20XX.
[7] Data-Driven Business Models: Challenges and Opportunities of Big Data Monica Bulger, Greg Taylor, Ralph Schroeder Oxford Internet Institute September 2014.
[8] A review of big data in health care: challenges and opportunities by Susan E White. This article was published in the following Dove press journal: Open Access Bioinformatics 31 December 2014.
[10] Challenges and Opportunities with Big Data by Alexandros Labrinidis University of Pittsburgh and H. V. Jagadish University of Michigan at The 38th International Conference on Very Large Data Bases, August 27th 31st 2012, Istanbul, Turkey. Proceedings of the VLDB Endowment, Vol. 5, No. 12
[11] Defining architecture components of the Big Data Ecosystem CONFERENCE PAPER · MAY 2014 DOI: 10.1109/CTS.2014.6867550 by Yuri Demchenko, Cees de Laat and Peter Membrey.
[12] 5 V's : Turning Big Data into Value. Available at: http://api.ning.com/files/sT-mI3qB-hcNTnTNHUCzaeKy84BTVrfBNizGWSBI-*BamQnzRvsIHNd594hfvtVlh*sdCfk6rBszB6pSByR-IjT4E9Zan*eR/bor55.PNG.
[13] Morley-Fletcher E. Big data healthcare: an overview of the challenges in data intensive healthcare. 2013. Available at http://ec.europa.eu/information_society/newsroom/cf/dae/ document.cfm?doc_id=3499 (accessed 17 August 2014)
[14] Patil HK, Seshadri R. Big data security and privacy issues in healthcare. In: 2014 IEEE International Congress on Big Data; 2014 June 27–July 2; Anchorage, AK, USA.
[15]Harsh Kishore Mishra Centre for Computer Science and Technology. Available at: http://www.slideshare.net/HarshMishra3/harsh-big-data-seminar-report?qid=a34637cb-fc80-49df-ac38-68f25ba4fe47&v=&b=&from_search=2
[16] Semantics-Empowered Approaches to Big Data Processing for Physical-Cyber-Social Applications by Krishnaprasad Thirunarayan and Amit Sheth in Semantics for Big Data AAAI Technical Report FS-13-04
[17] Addressing Bioinformatics Big Data Problems using Natural Language Processing: Help Advancing Scientific Discovery and Biomedical Research by EMDAD KHAN in Modern Computer Applications in Science and Education
[18] Big Data and Natural Language: Extracting Insight From Text in An Oracle White Paper September 2012
[19] Big Data is the Future of Healthcare in cognizant 20-20 insights Sep 2012.
[20] Big Data Analytics and its Application in Ecommerce by Uyoyo Zino Edosio in Conference Paper · April 2014