An Evolution of Big Data and its Challenges

Dr.P. Sujatha and Kanimozhi

Abstract--- Technological advancement such as cloud computing, internet of things, sensor web, sensor networks, video surveillance etc has drastically change the type and nature of data and its collection. Today data no longer implies simple text, but includes audio, video, images and animations. These are all extremely large and complex to store and manipulate as well as to make meaningful contribution in the application of data science. This paper explores an evolution of BIG DATA, a field of data science that deals with large complex data, storage and processing in order to derive quantifiable meaning and application.

Keywords--- Data Mining, Class, Association Rules, Clustering, Classification, Segmentation

I. INTRODUCTION

Data is becoming bigger and bigger every minutes, hour, daily, weekly, monthly, and yearly, not only due to the size of data, but also the nature of collected data (audio, video, images, animation). These are owing to the technological advancements in data collection techniques which has the capability of collecting environmental data (sensor based systems), surveillance data (video and audio), human physiological data (ECG, EEG). Besides it is necessary to derive more meaning (value) from collected data. For example common data like customer address, which a single organization may keep for contact and inventory purpose can be analyzed to extract several, meaning (value). Further it is used to determine locations where more orders were made with respect time (seconds, minutes, hours) and date (per day, week, month). These can further explain the factors behind such increase or decrease in order. Such information can help an organization to make real-time decisions (opening a new sales outlet in a location or not and also inventory management with respect to time date). Hence there is a need for advanced real-time methodology for analyzing, storing and presenting these data. Big Data is a determinant of the success of most technological advancement such as cloud computing, sensor web, internet of things, etc.

Quick Primer On Data Sizes

We take a quick prime on the data size measurements to confirm the exponential growth rate of data. Though this is relative to the type of data, for instance images, videos are larger than text and numerical data.

DATA EXPONENTIAL GROWTH CHART		
UNIT	**SIZE**	**Equals in Byte**
Bit (b)	1 or 0	
Byte(b)	8 bits	
Kilobyte(KB)	1,000 bytes	2^{10}bytes
Megabyte(MB)	1,000KB	2^{20} bytes
Gigabyte(GB)	1,000MB	2^{30} bytes
Terabyte (TB)	1,000GB	2^{40} bytes
Petabyte (PB)	1,000TB	2^{50} bytes
Exabyte(EB)	1,000PB	2^{60} bytes
Zettabyte(ZB)	1,000EB	2^{70} bytes
Yottabyte(YB)	1,000ZB	2^{80} bytes

II. LITERATURE SURVEY

In [1], Networked European Software and Services Initiative (NESSI) focused on the challenges accompanying the use of Big Data and counter measures to overcome them with the sole aim of opening up unexplored opportunities in Europe economy. More emphasis towards ensuring creating awareness of this new technology field and its enormous opportunities by polling more skilled professionals into this

Dr.P. Sujatha, Associate Professor, Department of Computer Science, Vels University, Chennai, India. E-mail:suja.research@gmail.com

Kanimozhi, Research Scholar, School of Computing Sciences, Vels University, Chennai, India. E-mail:kanimozhi572@gmail.com

domain as well as a conducive atmosphere with the right policy framework to make Europe more competitive against the rest of the world again. Sequel to this, they proposed to enhance research in Big Data, building an EU big Data eco system to enhance business.

Datasax [2] highlights the variation in interpretation of the term Big Data among different academia and industries. However more professionals doubt the feasibility of yielding any significant benefit using Big Data. They finally explored and presented a cross industrial advancements of Big Data and elaborated on the role of Big Data in modern competitive business environment.

ITU-T Technology Watch Report [3] examines the various application examples of Big Data as an emerging data science technology challenges and solutions in adoption of Big Data, as well as numerous similarities between them, also some enabling technology behind the emergence of this new data science technology field were outlined.

Gali Halevi, et al (4) examined the term Big Data with respect to literary view strictly, since they not only include numerous topics in Big Data, also includes challenges and solutions. Further they presented a draft on the emergence of Big Data from five points of view that are categories of published papers, preview on time/history, geographic and disciplinary output.

III. BIG DATA - CHARACTERISTICS

Big data is characterized by four Vs (volume, velocity, variety and veracity) briefly described below.

- **Volume:** [Data anytime, anywhere, by anyone and anything]

Volume is the only distinction between big data and other data analytics domain. It incorporates analysis of large varying datasets more effectively than traditional data mining. The size of the data is not specific but the nature of the dataset size is absolutely of concern. For example downloading three tweets per/seconds, One million HD movies per minute etc.

- **Velocity: [**Every millisecond counts]

The time taken from producing the input data to decide the output is a considered as a critical factor in the big data. New technologies should be capable of processing vast volumes of data in real time environment that increases the flexibility and hence organizations can respond to changes in the market as well as shifting customer preferences. Moreover big data systems should be capable of handling and linking data flows that are entering into the system at different frequencies.

- **Variety:** [The reality of data is messy]

Big data comprises various type and structure of data such as text, sensor data, call records, maps, audio, image, video, click streams, log files and more. Source data is dissimilar and hence this may take more time and effort to shape it into a required form which is fit for further processing and analysis. Data are generated when the human is interacted with the machine as well as when the machine is interacted with another machine. These data may be of different format. For example the Indian defense will collect video data, as well as other numerical data from sensors onboard to accurately identify and hit target. Industrial systems comprises of sensors that are monitoring the production line triggering actions which enable the smooth operation of the production line as well as monitor the state of each production plant automation parts and hence generating numerical data. Accordingly the format of datasets varies between source and application.

- **Veracity : [**Data in doubt]

Datasets were only generated and stored for archive purpose. Big data is all about extracting the meaning out of these archived data as well as the embedded benefits for future decision. Hence there is an immediate need for a system with the necessary features to segregate, analyze, and weigh numerous datasets in order to sustain veracity.

Characteristics	Description	Attributes
Volume	Based on the complete data analysis, data are generated, analyzed as well as managed.	• Exabyte • Zettabyte, • Yottabyte
Velocity	Rapidly how data is being created and changed as well as the speed at which data is transformed	• Batch • Near real-time • Real time • Streams
Variety	The degree of assortment of data from various sources both inside and outside of an organization.	• Degree of Structure • Complexity
Veracity	The quality and provenance of data.	• Consistency • Completeness • Integrity • Ambiguity

IV. CHALLENGES IN BIG DATA

To make Big data program a successful one, below five challenges has to be addressed first [7].

a) Uncertainty of the Data Management Landscape – There are many competing technologies, and within each technical area there are numerous rivals. Our first challenge is making the best choices while not introducing additional unknowns and risk to big data adoption.

b) The Big Data Talent Gap – The excitement around big data applications seems to imply that there is a broad community of experts available to help in implementation. However, this is not yet the case, and the talent gap poses our second challenge.

c) Getting Data into the Big Data Platform – The scale and variety of data to be absorbed into a big data environment can overwhelm the unprepared data practitioner, making data accessibility and integration our third challenge.

d) Synchronization Across the Data Sources – As more data sets from diverse sources are incorporated into an analytical platform, the potential for time lags to impact data currency and consistency becomes our fourth challenge.

e) Getting Useful Information out of the Big Data Platform – Lastly, using big data for different purposes ranging from storage augmentation to enabling high-performance analytics is impeded if the information cannot be adequately provisioned back within the other components of the enterprise information architecture, making big data syndication our fifth challenge.

V. CONCLUSION

We have entered an era of Big Data. 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. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data.

REFERENCE

[1] NESSI White Paper, 2012, “Big Data A New World of Opportunities”.

[2] Data: Beyond the Hype Why Big Data Matters to You, White Paper BY DATASTAX CORPORATION, October 2013

[3] “Big Data: Big today, normal tomorrow” , 2013, ITU-T Technology Watch Report

[4] Gali Halevi, Dr. Henk Moed , 2012, “The Evolution of Big Data as a Research and Scientific Topic- Overview of the Literature, Elserver research trends journal.

[5] Definitions of big data Big Gantz et al. (IDC definition); Understanding Big Data, Eaton et al. (IBM definition) ; The World According to LINQ, Meijer (Microsoft research)

[6] Challenges in Big data. Available online: https://www.progress.com/~/media/Progress/Documents/Papers/Addressing-Five-Emerging-Challenges- of-Big-Data.pdf