International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 2, 2017 ISSN: 2348-6600
Reference ID: IJCS-165 PAGE NO: 1004-1012
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

AN DETAILED SURVEY ON BIG DATA ANALYTICS IN TEXT AUDIO AND VIDEO

1.R.Briundha Devi 2.Dr.B.Radha
Assistant Professor, Department of Computer Science
Sree Saraswathi Thyagaraja College, Pollachi – 642107, TN, India
1.brindhadevirm@gmail.com
2. radhakbr10@gmail.com

ABSTRACT- this paper attempts to offer a broader definition of big data that captures its other unique and defining characteristics. Big data is the term for data sets so large and complicated that it becomes difficult to process using traditional data management tools or processing applications. This paper reveals most recent progress on big data and analytic methods used for big data. The potential value of big data analytics is great and is clearly established by a growing number of studies. There are keys to success with big data analytics, including a clear business need, strong committed sponsorship, alignment between the business and IT strategies, a fact-based decision making culture, a strong data infrastructure, the right analytical tools, and people skilled in the use of analytics. Main feature of this paper is its focus on analytics related to text audio and video big data.

Keywords— Big data analytics, Big data definition, unstructured data analytics.

I. INTRODUCTION

This paper documents the basic concepts relating to big data. It attempts to consolidate the hitherto fragmented discourse on what constitutes big data, what metrics define the size and other characteristics of big data, and what tools and technologies exist to harness the potential of big data. From corporate leaders to municipal planners and academics, big data are the subject of attention, and to some extent, fear. The sudden rise of big data has left many unprepared.

In the past new technological developments first appeared in technical and academic publications. The knowledge and synthesis later seeped into other avenues of knowledge mobilization, including books. The fast evolution of big data technologies and the ready acceptance of the concept by public and private sectors left little time for the discourse to develop and mature in the academic domain.

A key contribution of this paper is to bring forth the oft-neglected dimensions of big data. The popular discourse on big data, which is dominated and influenced by the marketing efforts of large software and hardware developers, focuses on predictive analytics and structured data. It ignores the largest component of big data, which is unstructured and is available as audio, images,

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 2, 2017 ISSN: 2348-6600
Reference ID: IJCS-165 PAGE NO: 1004-1012
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

video, and unstructured text. It is estimated that the analytics ready structured data forms only a small subset of big data. The unstructured data, especially data in video format, is the largest component of big data that is only partially archived. This paper is organized as follows. We begin the paper by defining big data. We highlight the fact that size is only one of several dimensions of big data. Other characteristics, such as the frequency with which data are generated, are equally important in defining big data. We then expand the discussion on various types of big data analytics, namely text, audio and video.

II. Defining Big Data

Laney(2001) suggested that Volume, Variety, and Velocity (or the Three V's) are the three dimensions of challenges in data management. The Three V's have emerged as a common framework to describe big data (Chen, Chiang, & Storey, 2012; Kwon, Lee, & Shin, 2014).For example, Gartner, Inc. defines big data in similar terms:

"Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."("Gartner IT Glossary, n.d.")

Similarly, Tech America Foundation defines big data as follows: "Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information." (TechAmericaFoundation's Federal Big Data Commission, 2012).

We describe the Three V's below.

Volume refers to the magnitude of data. Big data sizes are reported in multiple terabytes and petabytes. A survey conducted by IBM in mid-2012 revealed that just over half of the 1144 respondents considered data sets over one terabyte to be big data (Schroeck, Shockley, Smart, Romero-Morales, & Tufano, 2012).

One terabyte stores as much data as would fit on 1500 CDs or 220 DVDs, enough to store around 16 million Facebook photographs. Beaver, Kumar, Li, Sobel, and Vajgel (2010) report that Facebook processes up to one million photographs per second. One petabyte equals 1024 terabytes. Earlier estimates suggest that Facebook stored 260 billion photos using storage space of over 20 petabytes.

Definitions of big data volumes are relative and vary by factors, such as time and the type of data. What may be deemed big data today may not meet the threshold in the future because storage capacities will increase, allowing even bigger data sets to be captured. In addition, the type of data, discussed under variety , defines what is meant by 'big'. Two datasets of the same size may require different data management technologies based on their type, e.g., tabular versus video data. Thus, definitions of big data also depend upon the industry. These considerations there-fore make it impractical to define a specific threshold for big

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 2, 2017 ISSN: 2348-6600
Reference ID: IJCS-165 PAGE NO: 1004-1012
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

data volumes.

Variety refers to the structural heterogeneity in a dataset. Technological advances allow firms to use various types of structured, semi-structured, and unstructured data. Structured data, which constitutes only 5% of all existing data (Cukier, 2010), refers to the tabular data found in spread sheets or relational databases.

Text, images, audio, and video are examples of unstructured data, which sometimes lack the structural organization required by machines for analysis. Spanning a continuum between fully structured and unstructured data, the format of semi-structured data does not conform to strict standards. Extensible Markup Language (XML), a textual language for exchanging data on the Web, is a typical example of semi-structured data. XML documents contain user-defined data tags which make them machine-readable. A high level of variety, a defining characteristic of big data, is not necessarily new. Organizations have been hoarding unstructured data from internal sources (e.g., sensor data) and external sources (e.g., social media). However, the emergence of new data management technologies and analytics, which enable organizations to leverage data in their business processes, is the innovative aspect. For instance, facial recognition technologies empower the brick-and-mortar retailers to acquire intelligence about store traffic, the age or gender composition of their customers, and their in-store movement patterns.

This invaluable information is leveraged in decisions related to product promotions, placement, and staffing. Clickstream data provides a wealth of information about customer behaviour and browsing patterns to online retailers. Clickstream advises on the timing and sequence of pages viewed by a customer. Using big data analytics, even small and medium-sized enterprises (SMEs) can mine massive volumes of semi-structured data to improve website designs and implement effective cross-selling and personalized product recommendation systems.

Velocity refers to the rate at which data are generated and the speed at which it should be analysed and acted upon. The proliferation of digital devices such as smartphones and sensors has led to an unprecedented rate of data creation and is driving a growing need for real-time analytics and evidence-based planning. Even conventional retailers are generating high-frequency data. Wal-Mart, for instance, processes more than one million transactions per hour (Cukier, 2010).

The data emanating from mobile devices and flowing through mobile apps produces torrents of information that can be used to generate real-time, personalized offers for everyday customers. This data provides sound information about customers, such as geospatial location, demographics, and past buying patterns, which can be analyzed in real time to create real customer value. Given the soaring popularity of smartphones, retailers will soon have to deal with hundreds of thousands of streaming data sources that demand real time analytics. Traditional data management systems are not

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 2, 2017 ISSN: 2348-6600
Reference ID: IJCS-165 PAGE NO: 1004-1012
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

capable of handling huge data feeds instantaneously. This is where big data technologies come into play. They enable firms to create real time intelligence from high volumes of 'perishable' data.

In addition to the three V's, other dimensions of big data have also been mentioned. These include:

Veracity IBM coined Veracity as the fourth V, which represents the unreliability inherent in some sources of data. For example, customer sentiments in social media are uncertain in nature, since they entail human judgment. Yet they contain valuable information. Thus the need to deal with imprecise and uncertain data is another facet of big data, which is addressed using tools and analytics developed for management and mining of uncertain data.

Variability (and complexity) SAS introduced Variability and Complexity as two additional dimensions of big data. Variability refers to the variation in the data flow rates. Often, big data velocity is not consistent and has periodic peaks and troughs. Complexity refers to the fact that big data are generated through a myriad of sources. This imposes a critical challenge: the need to connect, match, cleanse and transform data received from different sources.

Value Oracle introduced Value as a defining attribute of big data. Based on Oracle's definition, big data are often characterized by relatively "low value density". That is, the data received in the original form usually has a low value relative to its volume. However, a high value can be obtained by analysing large volumes of such data.

III. BIG DATA ANALYTICS

The stored data does not generate business value, and this is true of traditional databases, data warehouses, and the new technologies for storing big data (e.g., Hadoop). Once the data is appropriately stored, however, it can be analyzed and this can create tremendous value. The overall process of extracting insights from big data can be broken down into five stages (Labrinidis &Jagadish, 2012), shown in Fig. 1. These five stages form the two main sub-processes: data management and analytics. Data management involves processes and supporting technologies to acquire and store data and to prepare and retrieve it for analysis. Analytics on the other hand, refers to techniques used to analyze and acquire intelligence from big data. Thus, big data analytics can be viewed as a sub-process in the overall process of 'insight extraction' from big data. In the following sections, we briefly review big data analytical techniques for structured and unstructured data.

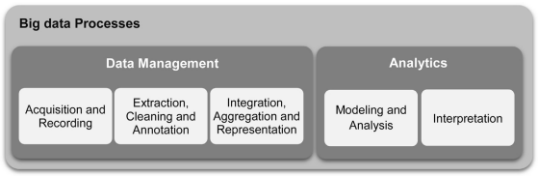

Fig. 3. Processes for extracting insights from big data.

3.1. Text Analytics

Text analytics (text mining) refers to techniques

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 2, 2017 ISSN: 2348-6600
Reference ID: IJCS-165 PAGE NO: 1004-1012
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

that extract information from textual data. Social network feeds, emails, blogs, online forums, survey responses, corporate documents, news, and call center logs are examples of textual data held by organizations. Text analytics involve statistical analysis, computational linguistics, and machine learning. Text analytics enable businesses to convert large volumes of human generated text into meaningfull summaries, which support evidence-based decision-making. For instance, text analytics can be used to predict stock market based on information extracted from financial news (Chung, 2014). We present a brief review of text analytics methods below.

Information extraction (IE) techniques extract structured data from unstructured text. For example, IE algorithms can extract structured information such as drug name, dosage, and frequency from medical prescriptions. Two sub-tasks in IE are Entity Recognition (ER) and Relation Extraction (RE) (Jiang, 2012). ER finds names in text and classifies them into predefined categories such as per-son, date, location, and organization. RE finds and extracts semantic relationships between entities (e.g., persons, organizations, drugs, genes, etc.) in the text. For example, given the sentence "Steve Jobs co-founded Apple Inc. in 1976", an RE system can extract relations such as Founder Of [Steve Jobs, Apple Inc.] or Founded In [AppleInc., 1976].

question answering (QA) techniques provide answers to ques-tions posed in natural language. Apple's Siri and IBM's Watson are examples of commercial QA systems. These systems have been implemented in healthcare, finance, marketing, and education. Similar to abstractive summarization, QA systems rely on complex NLP techniques. QA techniques are further classified into three categories: the information retrieval (IR)-based approach, the knowledge-based approach, and the hybrid approach.

IR-based QA systems often have three sub-components. First is the question processing, used to determine details, such as the question type, question focus, and the answer type, which are used to create a query. Second is document processing which is used to retrieve relevant pre-written passages from a set of existing documents using the query formulated in question processing. Third is answer processing, used to extract candidate answers from the output of the previous component, rank them, and return the highest-ranked candidate as the output of the QA system.

Sentiment analysis (opinion mining) techniques analyze opinionated text, which contains people's opinions toward entities such as products, organizations, individuals, and events. Businesses are increasingly capturing more data about their customers sentiments that has led to the proliferation of sentiment analysis (Liu, 2012). Marketing, finance, and the political and social sciences are the major application areas of sentiment analysis. Sentiment analysis techniques are further divided into three sub-groups, namely document-level, sentence-level, and aspect-based Document-level techniques determine whether the whole document expresses a negative or a positive sentiment. The assumption is that the document

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 2, 2017 ISSN: 2348-6600
Reference ID: IJCS-165 PAGE NO: 1004-1012
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

contains sentiments about a single entity. While certain techniques categorize a document into two classes,negative and positive, others incorporate more sentiment classes(like the Amazon's five-star system) (Feldman, 2013).

3.2. Audio analytics

Audio analytics analyze and extract information from unstructured audio data. When applied to human spoken language, audio analytics is also referred to as speech analytics. Since these techniques have mostly been applied to spoken audio, the terms audio analytics and speech analytics are often used interchangeably. Currently, customer call centers and healthcare are the primary application areas of audio analytics .Call centers use audio analytics for efficient analysis of thousands or even millions of hours of recorded calls. These techniques help improve customer experience, evaluate agent performance, enhance sales turnover rates, monitor compliance with different policies (e.g., privacy and security policies), gain insight into customer behavior, and identify product or service issues, among many other tasks. Audio analytics systems can be designed to analyze a live call, formulate cross/up-selling recommendations based on the customer's past and present interactions, and provide feedback to agents in real time. In addition, automated call centers use the Interactive Voice Response (IVR) platforms to identify and handle frustrated callers. In healthcare, audio analytics support diagnosis and treatment of certain medical conditions that affect the patient's communication patterns (e.g., depression, schizophrenia, and cancer) (Hirschberg, Hjalmarsson, & Elhadad, 2010). Also, audio analytics can help analyze an infant's cries, which contain information about the infant's health and emotional status (Patil, 2010).

Speech analytics follows two common technological approaches: the transcript-based approach (widely known as large-vocabulary continuous speech recognition, LVCSR) and the phonetic-based approach. These are explained below

.• LVCSR systems follow a two-phase process: indexing and searching. In the first phase, they attempt to transcribe the speech content of the audio. This is performed using automatic speech recognition (ASR) algorithms that match sounds to words. The words are identified based on a predefined dictionary. If the system fails to find the exact word in the dictionary, it returns the most similar one. The output of the system is a searchable index file that contains information about the sequence of the words spoken in the speech. In the second phase, standard text based methods are used to find the search term in the index file

.• Phonetic-based systems work with sounds or phonemes. Phonemes are the perceptually distinct units of sound in a specified language that distinguishes one word from another (e.g., the phonemes/k/and/b/differentiate the meanings of "cat" and "bat").Phonetic-based systems also consist of two phases: phonetic indexing and searching. In the first phase, the system translates the input speech into a sequence of phonemes. This

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 2, 2017 ISSN: 2348-6600
Reference ID: IJCS-165 PAGE NO: 1004-1012
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

is in contrast to LVCSR systems where the speech is converted into a sequence of words. In the second phase, the system searches the output of the first phase for the phonetic representation of the search terms.

3.3. Video analytics

Video analytics, also known as video content analysis (VCA),involves a variety of techniques to monitor, analyze, and extract meaningful information from video streams. Although video analytics is still in its infancy compared to other types of data mining (Panigrahi, Abraham, & Das, 2010), various techniques have already been developed for processing real-time as well as pre-recorded videos. The increasing prevalence of closed circuit television (CCTV) cameras and the booming popularity of video sharing web sites are the two leading contributors to the growth of computerized video analysis. A key challenge, however, is the sheer size of video data. To put this into perspective, one second of a high-definition video, in terms of size, is equivalent to over 2000 pages of text (Manyika et al., 2011). Now consider that 100 hours of video are uploaded to YouTube every minute (YouTube Statistics,n.d.). Big data technologies turn this challenge into opportunity. Obviating the need for cost-intensive and risk-prone manual processing, big data technologies can be leveraged to automatically sift through and draw intelligence from thousands of hours of video. As a result, the big data technology is the third factor that has contributed to the development of video analytics. The primary application of video analytics in recent years has been in automated security and surveillance systems. In addition to their high cost, labor-based surveillance systems tend to be less effective than automatic systems (e.g., Hakeem et al., 2012 report that security personnel cannot remain focused on surveillance tasks for more than 20 minutes).

The data generated by CCTV cameras in retail outlets can be extracted for business intelligence. Marketing and operations management are the primary application areas. For instance, smart algorithms can collect demographic information about customers, such as age, gender, and ethnicity. Similarly, retailers can count the number of customers, measure the time they stay in the store, detect their movement patterns, measure their dwell time in different areas, and monitor queues in real time. Valuable insights can be obtained by correlating this information with customer demographics to drive decisions for product placement, price, assortment optimization, promotion design, cross-selling, layout optimization, and staffing.

Another potential application of video analytics in retail lies in the study of buying behaviour of groups. Among family members who shop together, only one interacts with the store at the cash register, causing the traditional systems to miss data on buying patterns of other members. Video analytics can help retailers address this missed opportunity by providing information about the size of the group, the group's demographics, and the individual members' buying behaviour.

In terms of the system architecture, there exist

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 2, 2017 ISSN: 2348-6600
Reference ID: IJCS-165 PAGE NO: 1004-1012
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

two approaches to video analytics, namely server-based and edge-based:

•Server-based architecture: In this configuration, the video captured through each camera is routed back to a centralized and dedicated server that performs the video analytics. Due to bandwidth limits, the video generated by the source is usually compressed by reducing the frame rates and/or the image resolution. The resulting loss of information can affect the accuracy of the analysis. However, the server-based approach provides economies of scale and facilitates easier maintenance

.•Edge-based architecture: In this approach, analytics are applied at the 'edge' of the system. That is, the video analytics is performed locally and on the raw data captured by the camera. As a result, the entire content of the video stream is available for the analysis, enabling a more effective content analysis. Edge-based systems, however, are more costly to maintain and have a lower processing power compared to the server-based systems.

IV. CONCLUSIONS

In this paper we attempt to briefly review the concepts and analytics of big data in text audio and video. This review would be helpful to researchers to center of attention on the different analytics methods of big data. We found that big data is becoming increasingly common in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing normally use big data to measure the business values. So, data mining will be more and more useful in future.

REFERENCES

[1] Aggarwal, C. C. (2011). An introduction to social network data analytics. In C. C.Aggarwal (Ed.), Social network data analytics (pp. 1–15). United States: Springer.

Barbier, G., & Liu, H. (2011). Data mining in social media. In C. C. Aggarwal (Ed.),Social network data analytics (pp. 327–352). United States: Springer.

[2] Beaver, D., Kumar, S., Li, H. C., Sobel, J., & Vajgel, P. (2010). Finding a needle inhaystack: Facebook's photo storage. In Proceedings of the nineth USENIX con-ference on operating systems design and implementation (pp. 1–8). Berkeley, CA,USA: USENIX Association.

[3] S.Hameetha Begum, "Data Mining Tools and Trends – An Overview," International Journal of Emerging Research in Management &Technology, ISSN: 2278-9359

[4] Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS Quarterly, 36(4), 1165–1188.

[5] Chung, W. (2014). BizPro: Extracting and categorizing business intelligence fac-tors from textual news articles. International Journal of Information Management,34(2), 272–284.

[6] Cukier K., The Economist, Data, data everywhere: A special report on manag-ing information, 2010, February 25, Retrieved from http://www.economist.com/node/15557443.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 2, 2017 ISSN: 2348-6600
Reference ID: IJCS-165 PAGE NO: 1004-1012
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

[7] Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. National ScienceReview, 1(2), 293–314.
[8] Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. National ScienceReview, 1(2), 293–314.
[9] Feldman, R. (2013). Techniques and applications for sentiment analysis. Communi-cations of the ACM, 56(4), 82–89.
[10] Gundecha, P., & Liu, H. (2012). Mining social media: A brief introduction. Tutorials inOperations Research, 1(4)
[11] Jiang, J. (2012). Information extraction from text. In C. C. Aggarwal, & C. Zhai (Eds.),Mining text data (pp. 11–41). United States: Springer.
[12] Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data.Proceedings of the VLDB Endowment, 5(12), 2032–2033.
[13] Laney, D. (2001, February 6). 3-D data management: Controlling data volume, velocity and variety. Application Delivery Strategies by META Group Inc. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
[14] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–167.
[15] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135.
[16] Patil, H. A. (2010). "Cry baby": Using spectrographic analysis to assess neonatal health status from an infant's cry. In A. Neustein (Ed.), Advances in speech recog-nition (pp. 323–348). United States: Springer.
[17] Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., & Tufano, P. (2012).Analytics: The real-world use of big data. How innovative enterprises extractvalue from uncertain data. IBM Institute for Business Valu"
[18] Tang, L., & Liu, H. (2010). Community detection and mining in social media. SynthesisLectures on Data Mining and Knowledge Discovery, 2(1), 1–137.
[19] TechAmerica Foundation's Federal Big Data Commission. (2012). Demystifying bigdata: A practical guide to transforming the business of Government.
[20] YouTube Statistics (n.d.). Retrieved from http://www.youtube.com/yt/press/statistics.html