An Approach To Big Data Analytics And Its Significance

Ayesha Mariyam[a], Sk. Altaf Hussain Basha[b]

aDepartment of Computer Science and Engineering, Bhaskar Engineering College, Jntuh, India

b Department of Information Technology, Gokaraju Rangaraju Institute of Engineering and Technology, Jntuh, India

Corresponding Author : Ayesha Mariyam

Abstract

In recent years, the rapid development of Internet, Internet of Things and cloud computing have led to the explosive growth of data in almost every industry and business area Big data is a term that describes the large volume of data which is both structured and unstructured. Big data analytics is always characterized by 3V's.Although big data is a trending buzzword in both academic and industry. Its meaning is still covered by much conceptual indefiniteness. Big data analytics is the process of examining large datasets to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. Big data analytics helps organizations harness their data and use it identify new opportunities. That, in turn, leads to smarter business moves, more efficient operations, higher profits and happier customers. This paper presents a consolidated description of big data and analytical methods used for big data. It also presents the significance of big data and key research areas.

KEYWORDS- Big data, Analytics, Data Management

I. INTRODUCTION

Recent years have witnessed a dramatic increase in our ability to collect data from various devices, sensors in different formats, from independent or connected applications. This data has outpaced our capability to process, analyse, store and understand these datasets. Big data has now became a ubiquitous term in many parts of industry and academia. This paper concentrates on the basic concepts relating to big data, it describes what constitutes big data, what metrics define the size and other characteristics of big data.

This paper is organized as follows, the paper starts with the definition of big data. It highlights 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 importantin defining big data. The discussion on frequency with which data are generated, are equally important in defining big data. Then expand the discussion on various types of big data. Big data analytics helps organizations harness their data and use it to identifynew opportunities. That, in turn leads to smarter business moves, more efficient operations, higher profits and happier customer. Also the significance of big data analytics is discussed.

II. DEFINING BIG DATA

Data is growing at a huge speed making it difficult to handle such large amount of data (Exabyte's). Clearly size is the first characteristic that comes to mind considering the question "what is big data "and how data has to be qualify as 'big data'. However, other characteristics of big data have emerged recently. The three V's have emerged as a common framework to describe big data.

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."

Similarly, TechAmerica 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."

Doug Laney was the first in talking about 3V's in Big Data Management, as shown in fig1 below:

Volume: there is more data than ever before, its size continuously increasing, but not the percent of data that our tools can process.

Variety: there are many different types of data, as text, sensor data, audio, video, graph, and more.

Velocity: data is arriving continuously as streams of data, and we are interested in obtaining useful information from it in real time.

Nowadays, there are two more V's,

Variability: there are changes in the structure of data and how users want to interpret the data.

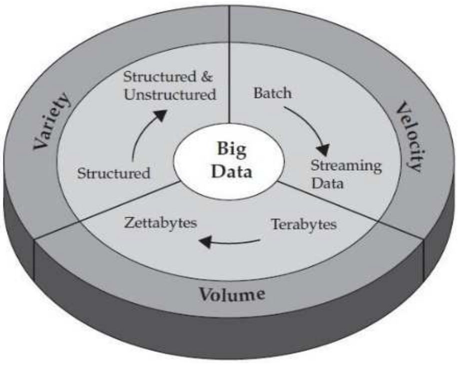

Fig1: Big Data, the 3V's

Value: business value that give organization a compelling advantage, due to the ability of making decisions based in answering questions that were previously considered beyond reach.

III. BIG DATA ANALYTICS

Big data are worthless in a vacuum. Its potential value is unlocked only when leveraged to drive decision making. To enable such evidence-based decision making, organizations need efficient processes to turn high volumes of fast moving and diverse data into meaningful insights. Big data analytics is the process of examining large datasets to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. The overall process of extracting insights from big data can be broken down into five stages (Labrinidis and Jagadish, 2012), shown in fig2. These five stages form the two main sub-process:

1. Data Management and
2. 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 analyse and acquire intelligence from big data. Thus, big data analytics can be viewed as a sub-process of ‘insight extraction’ from big data.

Thus, we briefly review big data analytical techniques for structured and unstructured data.

Thus, the following techniques represent a relevant subset of tools available for big data analytics.

i. Text analytics

Text analytics (text mining) refers to techniques that extract information from textual data. Social network feeds, emails, blogs, online forums, survey responses,

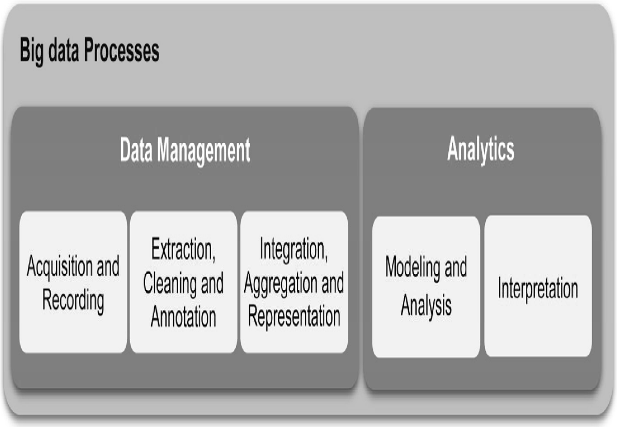
Big data Processes
Data Management
Analytics
Acquisition and Recording
Extraction, Cleaning and Annotation
Integration, Aggregation and Representation
Modeling and Analysis
Interpretation

Fig2: processes for extracting insights from big data.

Corporate documents, news, and call centre logs are examples of textual data held by organizations. Text analytics involve statistical analysis, computational linguists, and machine learning.

Text analytics enable businesses to convert large volumes of human generated text into meaningful summaries, which support evidence-based decision-making. 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.

Text summarization techniques automatically produce a succinct summary of a single or multiple documents. The resulting summary conveys the key information in the original text(s).Applications include scientific and news articles, advertisements, emails, and blogs.In order to parse the original text and generate the summary, abstractive summarization incorporates advanced Natural Language Processing (NLP) techniques. Question answering (QA) techniques provide answers to questions 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.

Sentiment analysis (opinion mining) techniques analyseopinionated text, which contains people’s opinions toward entities such as products, organizations, individuals, and events. Businesses are increasingly capturing more data about their customer’s 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.

ii. Audio analytics

Audio analytics analyse and extract information from unstructured audio data. When applied to human spoken language, audio analytics is also referred to as speech analytics. Currently, customer call centres and healthcare are the primary application areas of audio analytics.

iii. Video analytics

Video analytics, also known as video content analysis (VCA), involves a variety of techniques to monitor, analyse, and extract meaningful information from video streams. A key challenge, however, is the sheer size of video data. Now consider that 100 hours of video are uploaded to YouTube every minute. Video analytics can efficiently and effectively perform surveillance functions such as detecting breaches of restricted zones, identifying objects removed or left unattended, detecting loitering in a specific area, recognizing suspicious activities, and detecting camera tampering.

iv. Social media analytics

Social media analytics refer to the analysis of structured and unstructured data from social media channels. Social media is abroad term encompassing a variety of online platforms that allow users to create and exchange content. Social media can be categorized into the following types: Social networks (e.g., Facebook and LinkedIn), blogs (e.g., Blogger and WordPress), microblogs (e.g. Twitter and Tumblr), social news (e.g., Digg and Reddit), social bookmarking (e.g., Delicious and Stumble Upon), media sharing(e.g., Instagram and YouTube), wikis (e.g., Wikipedia and Wikihow),question-and-answer sites (e.g., Yahoo! Answers and Ask.com) and review sites (e.g., Yelp, TripAdvisor) (Barbier& Liu, 2011; Gundecha& Liu, 2012).

v. Predictive analytics

Predictive analytics comprise a variety of techniques that predict future outcomes based on historical and current data. Inpractice, predictive analytics can be applied to almost all disciplines from predicting the failure of jet engines based on the stream ofdata from several thousand sensors, to predicting customers' nextmoves based on what they buy, when they buy, and even what theysay on social media.

IV. SIGNIFICANCE OF BIG DATA

Due to its great value, big data has been essentially changing and transforming the way we live, work, and think[4]. In what follows, we describe in detail the significance of big data in various perspectives.

i. Significance to national development

At present, the world has completely entered the era of the information age. The extensive use of Internet, Internet of Things, Cloud Computing, and other emerging IT technologies has made various data sources increasing at an unprecedented rate, while making the structures and types of data increasingly complex. Depth analysis and utilization of big data will play an important role in promoting sustained economic growth of countries and enhance the competitiveness of companies.

In the future, big data will become a new point of economic growth. With big data, companies will upgrade and transform to the mode of Analysis as a Service (AaaS), thereby changing the ecology of the IT and other industries. In this context, the global giants of the IT industry (such as IBM, Google, Microsoft, and Oracle) have already begun their technical development planning in the big data era.

At the national level, the capacity of accumulating, processing, and utilizing vast amounts of data will become a new landmark of a country's strength. The data sovereignty of a country in cyberspace will be another great power-game space besides land, sea, air, and outer spaces.

In general, the Western countries, represented by the United States, are moving under their national agenda towards a modernization of their national strength through big data research and applications. It is anticipated that future economic and political competitions among countries will be based on exploiting the potential of big data, among other traditional aspects.

.

ii. Significance to industrial upgrades

Big data is currently a common problem faced by many industries, and it brings grand challenges to these industries digitization and information. Research on common problems of big data, especially on breakthroughs of core technologies, will enable industries to harness the complexity induced by data interconnection and to master uncertainties caused by redundancy and/or shortage of data. Everyone hopes to mine from big data demand-driven information, knowledge and even intelligence and ultimately taking full advantage of the big value of big data. This means that data is no longer a byproduct of the industrial sector, but has become a key nexus of all aspects. In this sense, the study of common problems and core technologies of big data will be the focus of the new generation of IT and its applications. It will not only be the new engine to sustain the high growth of the information industry, but also the new tool for industries to improve their competitiveness.

For example, in recent years, cloud computing has rapidly evolved from a vague concept in the beginning to a mature hot technology. Many big companies, including Google, Microsoft, Amazon, Facebook, Alibaba,2Baidu,3Tencent,4and other IT giants, are working on cloud computing technologies and cloud-based computing services. Big data and cloud computingis seen as two sides of a coin: big data is a killer application of cloud computing, whereas cloud computing provides the IT infrastructure to big data. The tightly coupled big data and cloud computing nexus are expected to change the ecosystem of Internet, and even affect the pattern of the entire information industry.

iii. Significance to scientific research

Big data has caused the scientific community to re-examine its methodology of scientific research[5]and has triggered a revolution in scientific thinking and methods.

The emergence of big data has spawned a new research paradigm; that is, with big data, researchers may only need to find or mine from it the required information, knowledge and intelligence. They even do not need to directly access the objects to be studied. In 2007, the late Turing Award winner, Jim Gray, depicted in his last speech the fourth paradigm of data-intensive scientific research[5], which separates data-intensive science from computational science. Gray believed that the fourth paradigm may be the only systemic way for solving some of the toughest global challenges we face today. In essence, the fourth paradigm is not only a change in the way of scientific research, but also a change in the way that people think.

iv. Significance to emerging interdisciplinary research

Big data technologies and the corresponding fundamental re-search have become a research focus in academia. An emerging interdisciplinary discipline called data science has been gradually coming into place. This takes big data as its research object and aims at generalizing the extraction of knowledge from data. It spans across many disciplines, including information science, mathematics, social science, network science, system science, psychology, and economics. It employs various techniques and theories from many fields, including signal processing, probability theory, machine learning, statistical learning, computer programming, data engineering, pattern recognition, visualization, uncertainty modeling, data warehousing, and high performance computing.

Many research centers/institutes on big data have been established in recent years in different universities throughout the world (such as the University of California at Berkeley, Columbia University, New York University, Tsinghua University, Eindhoven University of Technology, and Chinese University of Hong Kong). Lots of universities and research institutes have even set up under-graduate and/or postgraduate courses on data analytics for cultivating talents, including data scientists and data engineers.

v. Significance to helping people better perceive the present

Big Data, especially big networked data, contains a wealth of societal information and can thus be viewed as a network mapped to society. To this end, analyzing big data and further summarizing and finding clues and laws it implicitly contains can help us better perceive the present. Deep mining information contained in big data can also help people make better decisions. For example, in the presidential election of the United States in November 2012, Barack Obama's campaign team helped Obama by analyzing big data in order to beat Romney and to get re-elected.5In the eighteen months be-fore Election Day, Obama's data analysis team created a huge data processing system. Through real-time data collection and analysis, not only could it tell the campaign team how to find voters and to get their attention, but it also analyzed the tendency for voters to vote. Every night, the data analysis team conducted simulation on the election and presented simulation results in the next day to help understand the possibility that Obama might win in some areas, based on which the team can allocate resources more precisely. Later facts demonstrated that the data analysis team played a crucial role in Obama's reelection, far beyond people's imagination.

vi. Significance to helping people better predict the future

Through effective integration and accurate analysis on multi-source heterogeneous big data, better predictions of future trends of events can be achieved. It is possible for big data analysis to even promote sustainable developments of society and economy and further give birth to new industries related to data services.

The ability of big network data has been being highly developed and effectively applied in the field of security and military.

Big data-based predictive analysis has been applied to address societal issues, including public health and economic development. Ginsberg, *et al.*found that, if the volume of queries submitted to Google and with keywords like "flu symptom" and "flu treatment" increase in a region, then after a few weeks, the number of influenza patients to the emergency rooms of hospitals in the corresponding area will increase accordingly. With this discovery, they will be able to predict outbreaks of influenza and deploy countermeasures in advance.

vii. Significance to higher education

The use of Big Data and analytics in higher education is relatively new practice. It is also one of the future areas of research. The Horizon reports of 2011 and 2012 indicate the inclusion of analytics in higher education in span of three to five years. Analytics projects to be successful in an institution would require data, technology, statistical requirements and above all skill and leadership.

Any effort for planning and implementation of an analytics project in an institution would require leaders committed to decision making based on the institutional data. The role to be played by analytics relies on the institutions vision of the next generation learning system. To adapt analytics an institution should identify leaders who can use data to solve complex issues.

- Identify the key values on which data can be measured,
- Identify tools and models suitable to their requirements.
- Embed analytics in institutional process.
- Instill a plan for effective communication.

viii. Significance to stock market

The Stock market process is full of uncertainty and is affected by many factors. Hence the Stock market prediction is one of the important exertions in finance and business. There are two types of analysis possible for prediction, technical and fundamental. In this paper both technical and fundamental analysis are considered. Technical analysis is done using historical data of stock prices by applying machine learning and fundamental analysis is done using social media data by applying sentiment analysis.

Social media data has high impact today than ever, it can aide in predicting the trend of the stock market. The method involves collecting news and social media data and extracting sentiments expressed by individual. Then the correlation between the sentiments and the stock values is analyzed.

The learned model can then be used to make future predictions about stock values. It can be shown that this method is able to predict the sentiment and the stock performance and its recent news and social data are also closely correlated.

V. CONCLUSION

Big data has made a strong impact in almost every sector and industry today

The focus of this paper is to define, understand, and what makes data a big data. The paper first defined what is big data to consolidate the divergent discourse on big data. Various definitions of big data are presented, whichtells the fact that size is only one measure of big data. Other measures, such as velocity and variety are also have significance. The paper’s primary focus has been on analyticsto gain valid and valuable insights from big data. We reviewed analyticstechniques for text, audio, video, and social media data, as well aspredictive analytics. Big data has made a strong impact in almost every sector and industry today. In this paper, we have briefly reviewed the significance of big data. Various machine learning techniques are required for analytics which are not mentioned and is beyond the scope of this paper.

REFERENCES

[1] Labrinidis and Jagadish 2012, A. Labrinidis and H. Jagadish, Challenges and Opportunities with Big Data, In *Proc. of the VLDB Endowment*, 5(12):2032-2033, 2012.

[2] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–167.

[3] Barbieri, G., & Liu, H. (2011). Data mining in social media. In C. C. Aggarwal (Ed.), Social network data analytics (pp. 327–352). United States: Springer. Beaver, D., Kumar, S., Li, H. C., Sobel, J., & Vajgel, P. (2010). Finding a needle.

[4] V. Mayer-Schonberger, K. Cu kier, Big Data: A Revolution That Will Transform How We Live, Work, and Think, Houghton Mifflin Harcourt, 2013.

[5] T. Hey, S. Tansey, K. Tolle (Eds.), the Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Corporation, 2009.

[6] J. Gantz and D. Remsel, ‘‘The digital universe in 2020: Big data, bigger digital shadows and biggest growth in the far east,’’ in Proc. IDC iView, IDC Anal. Future, 2012.

[7] McKinsey, ‘‘Big data: The next frontier for innovation, competition, productivity,’’ in McKinsey Global Institute Report, 2011.