IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 8, August 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008188|

Social Network Analysis using BIGDATA

Dr. Sowmya Naik P T, Dr. Manjunath R, Dr. K.N. Narasimha Murthy[3], Ambika P.R[4]

Professor, Dept. of CSE., RR Institute f Technology, Visvesvaraya Technological University, Bangalore, India

Professor, Dept. of CSE., City Engineering College, Visveswaraya Technological University, Bangalore, India

Professor, Dept. of CSE., Christ University, Bangalore, India

Assistant Professor, Dept. of CSE., City Engineering College, Visvesvaraya Technological University, Bangalore, India

ABSTRACT: BigData refers to "A Growing Torrent"- the massive amounts and varieties of information, particularly in unstructured form being generated by websites, sensors, social media and others. Big Data recently has become a new ubiquitous term to describe large datasets. The main challenge in handling BigData lies in capturing data, providing a proper representation, managing and analysing data to extract meaningful information in the given domain in a fast and efficient manner. Effective management and analysis of Big Data would bring great benefits and unique opportunities to the users. Big data analytics is a technology-enabled strategy for gaining richer, deeper, and more accurate insights into customers, partners and the business and ultimately gaining competitive advantage. By processing a steady stream of real-time data, organizations can make time-sensitive decisions faster than ever before, monitor emerging trends, course-correct rapidly and jump on new business opportunities. In this paper, we present a brief history of BigData and a life cycle of BigData process model. In this paper , we have investigated many data analytics methods and finally we introduced the *Next Generation Analytics.*

KEYWORDS: Bigdata, Bigdata analytics, Social network analysis.

I. INTRODUCTION

"Big Data" is a term encompassing the use of techniques to capture, process, analyze and visualize potentially large datasets in a reasonable timeframe not accessible to standard IT technologies. By extension, the platform, tools and software used for this purpose are collectively called "Big Data technologies". Big data research is a vast field that connects with many enabling technologies. The new approach redefines the way data is managed and analyzed by leveraging the power of a distributed grid of computing resources. New technologies are emerging to make unstructured data analytics possible and cost-efficient.

1.1 CHARACTERISTICS OF BIGDATA

In this section, BigData characteristics are described using 3v's(Volume, Variety, Velocity) concept.

Volume: Big data implies enormous volumes of data. It used to be employees created data. Now that data is generated by machines, networks, sensors and human interaction on systems like social media the volume of data to be analyzed is massive.

Variety: Variety refers to many types of data both structured and unstructured. We used to store data from sources like spreadsheets and databases. Now data comes in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. This variety of unstructured data creates problems for storage, mining and analyzing data.

Velocity: Big Data Velocity deals with the pace at which data flows in from sources like business processes, machines, networks and human interaction with things like social media sites, mobile devices, etc. The flow of data is massive and continuous.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008188|

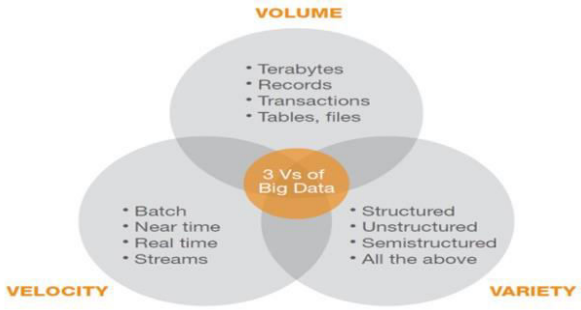

Fig 1. 3V's Of BigData

1.2 A BRIEF HISTORY OF BIGDATA

‘‘Big data’’ is designated to mean large, diverse, and complex datasets that are generated from various data sources. To understanding the history of big data, i.e., how it evolved into its current stage. In this paper, the history of big data is presented in terms of the data size of interest. Under this framework, the history of big data is tied tightly to the capability of efficiently storing and managing larger and larger datasets, with size limitations expanding by orders of magnitude as shown in Figure.1.2.

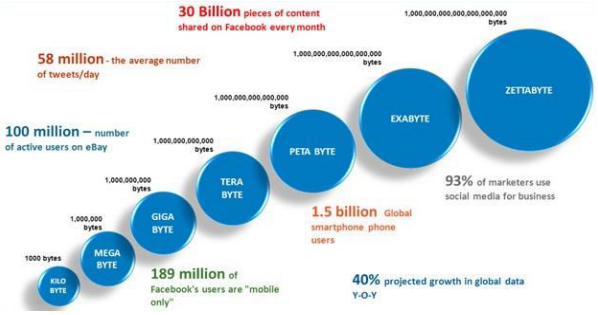

Fig 2. A Brief history of BigData

The history of big data can be roughly split into the following stages:

- *Megabyte to Gigabyte:* In the 1970s and 1980s, historical business data introduced the earliest ‘‘big data’’ challenge in moving from megabyte to gigabyte sizes. The urgent need at that time was to house that data and run relational queries for business analyses and reporting. Research efforts were made to give birth to the ‘‘database machine’’ that featured integrated hardware and software to solve problems. After a period of time, it became clear that hardware-specialized database machines could not keep pace with the progress of general-purpose computers. Thus, the descendant database systems are software systems that impose few constraints on hardware and can run on general-purpose computers.

- *Gigabyte to Terabyte:* In the late 1980s, the popularization of digital technology caused data volumes to expand to several gigabytes or even a terabyte, which is beyond the storage and/or processing capabilities of a single large computer system. Data parallelization was proposed to extend storage capabilities and to improve performance by distributing data and related tasks, such as building indexes and evaluating queries, into disparate hardware.

- *Terabyte to Petabyte :* During the late 1990s, when the database community was admiring its ‘‘finished’’ work on the parallel database, the rapid development of Web 1.0 led the whole world into the Internet era, along with massive semi-structured or unstructured web-pages holding terabytes or petabytes (PBs) of data. The resulting need for search companies was to index and query the mushrooming content of the web. Unfortunately, although parallel databases handle structured data well, they provide little support for unstructured data. Additionally, systems capabilities were limited to less than several terabytes. To address the challenge of web-scale data management and analysis, Google created Google File System (GFS)[5]and MapReduce [4] programming model.

- *Petabyte to Exabyte*: Under current development trends, data stored and analyzed by big companies will

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008188|

undoubtedly reach the PB to exabyte magnitude soon. However, current technology still handles terabyte to PB data; there has been no revolutionary technology developed to cope with larger datasets. In Jun. 2011, EMC published a report entitled ‘‘Extracting Value from Chaos’’ [6].

1.3 BIG-DATA PARADIGMS: STREAMING VS. BATCH

Big data analytics is the process of using analysis algorithms running on powerful supporting platforms to uncover potentials concealed in big data, such as hidden patterns or unknown correlations. According to the processing time requirement, big data analytics can be categorized into two alternative paradigms:

•*Streaming Processing:*The start point for the streaming processing paradigm [7] is the assumption that the potential value of data depends on data freshness. Thus, the streaming processing paradigm analyzes data as soon as possible to derive its results.

•*Batch Processing:*In the batch-processing paradigm,data are first stored and then analyzed. MapReduce [4]has become the dominant batch-processing model. The core idea of MapReduce is that data are first divided into small chunks. Next, these chunks are processed in parallel and in a distributed manner to generate intermediate results. The final result is derived by aggregating all the intermediate results. This model schedules computation resources close to data location, which avoids the communication overhead of data transmission. There are many differences between these two processing paradigms, as summarized in Table 1. Moreover, some research effort has been made to integrate the advantages of these two paradigms.

Table1. Stream processing Vs Batch processing

	Streaming processing	Batch processing
Input	Stream of new data or updates	Data chunks
Data Size	Infinite or unknown	Known and Finite
Storage	Not store	store
Hardware	Limited amount of memory	Multiple CPUs, memories
Time	A few seconds or Milliseconds	Much longer
Processing	A single or few passes over data	Processed in multiple passes
Application	Sensor networks, web mining, Traffic monitoring	Widely adopted in almost every domain

II. BIGDATA PROCESS MODEL

In this section, we focus on the process model for big data analytics. This model consists of four stages (generation, acquisition, storage, and Analytics). A big-data system is complex, providing functions to deal with different phases in the digital data life cycle, ranging from its birth to its destruction as shown in the figure 2. The details for each phase are explained as follows.

1.*Data generation:* concerns how data are generated. In this case, the term ‘‘big data’’ is designated to mean large, diverse, and complex data sets that are generated from various longitudinal and/or distributed data sources, including sensors, video, and other available digital sources. However, there are overwhelming technical challenges in collecting, processing, and analyzing these datasets that demand new solutions to embrace the latest advances in the information and communications technology domain.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008188|

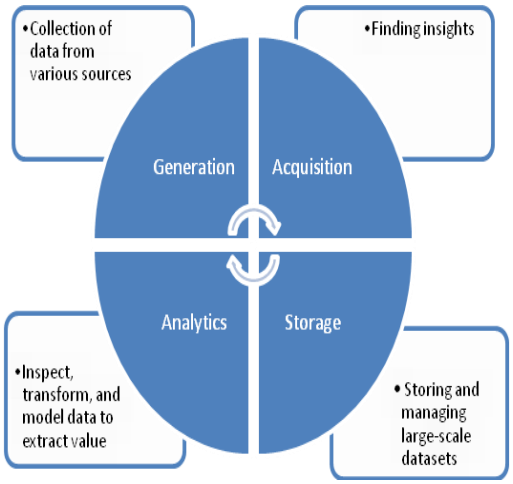

Fig 3. Big data Process model

2. *Data acquisition:* refers to the process of obtaining information and is subdivided into data collection, data transmission, and data pre-processing. First, because data may come from a diverse set of sources, websites that host formatted text, images and/or videos - data collection refers to dedicated data collection technology that acquires raw data from a specific data production environment. Second, after collecting raw data, we need a high-speed transmission mechanism to transmit the data into the proper storage sustaining system for various types of analytical applications. Finally, collected datasets might contain many meaningless data, which unnecessarily increases the amount of storage space and affects the consequent data analysis. For instance, redundancy is common in most datasets collected from sensors deployed to monitor the environment, and we can use data compression technology to address this issue. Thus, we must perform data pre-processing operations for efficient storage and mining.

3.*Data storage:* concerns persistently storing and managing large-scale datasets. A data storage system can be divided into two parts: hardware infrastructure and data management. The hardware infrastructure should be able to scale up and out and be able to be dynamically reconfigured to address different types of application environments. Data management software is deployed on top of the hardware infrastructure to maintain large-scale datasets. Additionally, to analyze or interact with the stored data, storage systems must provide several interface functions, fast querying and other programming models.

4.*Data analysis:* leverages analytical methods or tools to inspect, transform, and model data to extract value. Many application fields leverage opportunities presented by abundant data and domain-specific analytical methods to derive the intended impact. Although various fields pose different application requirements and data characteristics, a few of these fields may leverage similar underlying technologies. Emerging analytics research can be classified into six critical technical areas: structured data analytics, text analytics, multimedia analytics, web analytics, network analytics, and mobile analytics. This classification is intended to highlight the key data characteristics of each area.

III. METHODOLOGY

TAXONONOMY OF BIGDATA ANALYTICS

The most important stage of the big data process model is data analysis, the goal of which is to extract useful values, suggest conclusions and/or support decision-making.

PURPOSE AND CATEGORIES

Data analytics addresses information obtained through observation, measurement, or experiments about a phenomenon of interest. The aim of data analytics is to extract as much information as possible that is pertinent to the subject under consideration. The nature of the subject and the purpose may vary greatly. The following lists only a few potential purposes:

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008188|

• To extrapolate and interpret the data and determine how to use it,
• To check whether the data are legitimate,
• To give advice and assist decision-making,
• To diagnose and infer reasons for fault, and
• To predict what will occur in the future.

In this section we discuss six types of big data application, organized by data type: structured data analytics, text analytics, web analytics, multimedia analytics, network analytics, and mobile analytics.

A. STRUCTURED DATA ANALYTICS

A large quantity of structured data is generated in the business and scientific research fields. Management of these structured data relies on the mature RDBMS, data warehousing, OLAP, and BPM [2]. Data analytics is largely grounded in data mining and statistical analysis, as described above. These two fields have been thoroughly studied in the past three decades. Recently, deep learning, a set of machine-learning methods based on learning representations, is becoming an active research field. Most current machine-learning algorithms depend on human-designed representations and input features, which is a complex task for various applications.

B. TEXT ANALYTICS

Text is one of the most common forms of stored information and includes e-mail communication, corporate documents, webpages, and social media content. Hence, text analytics is believed to have higher commercial potential than structured data mining. In general, text analytics, also known as text mining, refers to the process of extracting useful information and knowledge from unstructured text. Text mining is an interdisciplinary field at the intersection of information retrieval, machine learning, statistics, computational linguistics, and, in particular, data mining. Most text mining systems are based on text representation and natural language processing (NLP), with emphasis on the latter. There are several technologies have been developed for text mining, including information extraction, topic modelling, summarization, categorization, clustering, question answering, and opinion mining. Information extraction refers to the automatic extraction of specific types of structured information from text.

C. WEB ANALYTICS

Over the past decade, we have witnessed an explosive growth of web pages, whose analysis has emerged as an active field. Web analytics aims to retrieve, extract, and evaluate information for knowledge discovery from web documents and services automatically. This field is built on several research areas, including databases, information retrieval, NLP, and text mining. We can categorize web analytics into three areas of interest based on which part of the web is mined: web content mining, web structure mining, and web usage mining [3]. Web content mining is the discovery of useful information or knowledge from website content. However, web content may involve several types of data, such as text, image, audio, video, symbolic, metadata, and hyperlinks. Because most of the web content data are unstructured text data, much of the research effort is centered on text and hypertext content. Hypertext mining involves mining semi-structured HTML pages that have hyperlinks. Supervised learning or classification plays a key role in hypertext mining, such as in email management, newsgroup management, and maintaining Web structure mining is the discovery of the model underlying link structures on the web. Here, structure represents the graph of links in a site or between sites. The model is based on the topology of the hyperlink with or without link description.

D. MULTIMEDIA ANALYTICS

Recently, multimedia data, including image, audio, and video, has grown at a phenomenal rate and is almost ubiquitous. Multimedia content analytics refers to extracting interesting knowledge and understanding the semantics captured in multimedia data. Because multimedia data are diverse and more information-rich than the simple structured data and text data in most of the domains, information extraction involves overcoming the semantic gap of multimedia data. The research in multimedia analytics covers a wide spectrum of subjects, including multimedia summarization, multimedia annotation, multimedia indexing and retrieval, multimedia recommendation, and multimedia event detection, to name only a few recent areas of focus. Audio summarization can be performed by simply extracting salient words or sentences from the original data. Video summarization involves synthesizing the most important or representative sequences of the video content and can be static or dynamic.

E. NETWORK ANALYTICS

The social media explosion has seen consumers publishing their lives online via tweets, photo uploads, “likes”, status updates, and so on. Much of this material is unstructured, has limited relevance to brands and exists in vast quantities,

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008188|

making it challenging to get value from the data. In working with social media, many organizations are encountering these big data challenges for the first time. Typically, social networks contain a tremendous amount of linkage and content. Social analytics describes the process of measuring, analyzing and interpreting the results of interactions and associations among people, topics and ideas. Social network [or media] analysis involves collecting data from multiple sources, identifying relationships, and evaluating the impact, quality or effectiveness of a relationship. Web Analytics and Social Analytics will be used together for some time to come, but the current focus is mainly on Social Analytics.

IV. BIG SOCIAL:NEXT-GENERATION ANALYTICS

BigData Analytics is becoming possible to run simulations or models to predict the future outcome, rather than to simply provide backward-looking data about past interactions, and to do these predictions in real-time to support each individual business action. The social media explosion has seen consumers publishing their lives online via tweets, photo uploads, “likes”, status updates, and so on. Much of this material is unstructured, has limited relevance to brands and exists in vast quantities, making it challenging to get value from the data. In working with social media, many organizations are encountering these big data challenges. The interesting thing is that, in these times of Big Data, Web and Social Analytics are evolving to a further stage which refers as " Next-Generation Analytics" Fig 4.

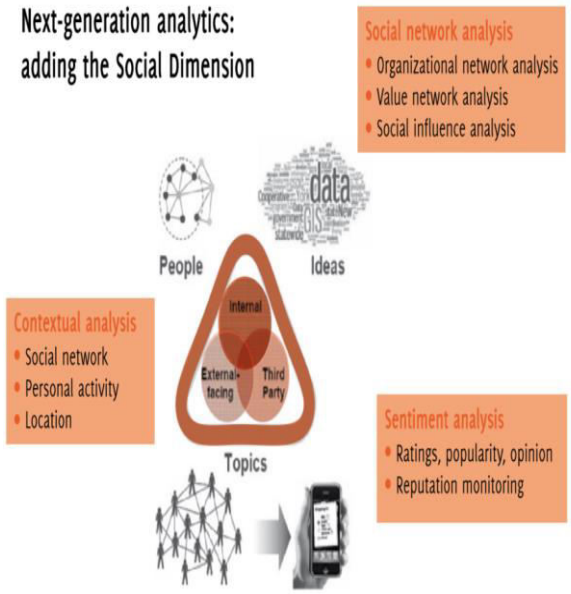

Figure 4. Next-generation analytics: adding the social Dimension.

In next generation analytics, "social" network analysis, sentiment analysis and contextual analysis plays a very important role. Social networking has a lot of positive influence on society in today‘s world, ranging from academics, parenting, industry, and to other social and relational part of human life. Every individual will one day like to know in the future a family tree from which he belongs based on dataset stored on his generation years before and currently. The social networks provide data for such opportunity in the future[9]. Twitter is an online social networking service and micro blogging service that enables its users to send and read text-based messages of up to 140 characters, known as "tweets". Public tweets of the Twitter are taken as the Big Data source for the sentiment analysis[10].

Big Data is Improving Decisions in Most Companies[11]

Slicing and dicing huge volumes and varieties of digital data can keep data scientists busy for days or even weeks. But if the insights they derive do not provide useful guidance – or if business managers do not utilize that guidance – all that spending is futile. To find out, the survey first asked participants whether their initiatives had improved decision making in the business. The answer for the clear majority – 80% -- was indeed yes. The lowest percentage of improvement in decision making was in the U.S. (77%) while the highest was in Latin America (86%). (See Fig 5.)[11]

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008188|

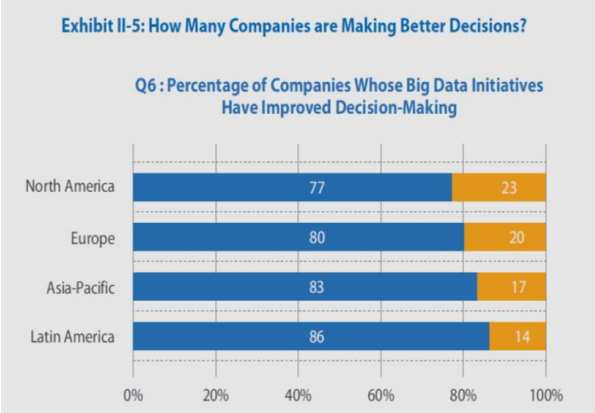

Figure 5. Percentage of companies whose big data initiatives have improved decision making[11]

- The company has a Big Data initiative(s) in place, and it has improved decision-making in the business.
- The company has a Big Data initiative(s) in place, and it hasn‟t yet improved decision-making in the business.

V. CONCLUSION

The era of big data is upon us, bringing with it an urgent need for advanced data acquisition, management, and analysis mechanisms. In this paper, we have presented the concept of big data, a BigData paradigm and highlighted the big process model, which covers the big data lifecycle. The big data process model consists of four phases: data generation, data acquisition, data storage, and data analysis. we present a survey on different categories of BigData analytics and journey towards the next generation analytics. This paper presents how BigData initiatives have improved the decision making of companies.

REFERENCES

[1] S. K. Pal, V. Talwar, and P. Mitra, ‘‘Web mining in soft computing ramework: Relevance, state of the art and future directions,’’ IEEE Trans. Neural Netw., vol. 13, no. 5, pp. 1163–1177, 2002.

[2] S. Chaudhuri, U. Dayal, and V. Narasayya, ‘‘An overview of business intelligence technology,’’ Commun. ACM, vol. 54, no. 8, pp. 88–98, 2011.

[3] J. Gantz and D. Reinsel, ‘‘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.

[4]J.DeanandS.Ghemawat, ‘‘Mapreduce:Simplifieddata processing on large clusters,’’ Commun. ACM, vol. 51, no. 1, pp. 107–113, 2008.

[5] S. Ghemawat, H. Gobioff, and S.-T. Leung, ‘‘The Google file system,’’ in Proc. 19th ACM Symp. Operating Syst. Principles, 2003, pp. 29–43.

[6] J. Gantz and D. Reinsel, ‘‘Extracting value from chaos,’’ in Proc. IDC iView, 2011, pp. 1–12.

[7] N. Tatbul, ‘‘Streaming data integration: Challenges and opportunities,’’ in Proc. IEEE 26th Int. Conf. Data Eng. Workshops (ICDEW), Mar. 2010, pp. 155–158.

[8] Naushar Khan et. Al, ‘‘Big data:survey, technologies, opportunities and challenges”, Hindawi Publishing Corporation Scientific World Journal Volume 2014, pp. 10-15.

[9] Quist-Aphetsi Kester," Optimization of Searches on Social Networks using Graph theory and Formal Concepts Analysis", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 1, January 2013,pp.1-2.

[10] Vibhavari Chavan et al, "Survey on Big Data", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 7932-7939, pp.1-4.

[10] T.K.Das1 , P.Mohan Kumar2, " BIG Data Analytics: A Framework for Unstructured Data Analysis", International Journal of Engineering and Technology (IJET), Vol 5 No 1 Feb-Mar 2013.

[11] The Emerging Big returns on big data,TCS-Big-Data-Global-Trend-Study-2013,mar 21,2013.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details