BIG DATA ANALYTICS – TYPES, BENEFITS, AND BARRIERS

[1]Hardeep Singh, [2]Sandeep Kaur, [3]Satveer Kour
[1]Assistant Professor, [2]Assistant Professor, [3]Assistant Professor
Department of Computer Science & Engineering,
GNDU RC, Sathiala-143205, Amritsar, Punjab
E-mail: [1]hardeepsinghcse12@gmail.com, [2]sandeepkaurcse@gmail.com, [3]rattansatveer1985@gmail.com

Abstract: *Big data is the current state of the art topic creating its unique place in the research and industry minds to look into depth of topic to get valuable results needed to meet the future data mining and analysis needs. Big data refers to the fast moving, large size of different structural form data increasing at fast pace. So, there also prevail the need of to analyze the data to get valuable data results from it . This paper deals with analytic emphasis on big data and why it is needed. In this paper, different sections through an overlook on different aspects on big data such as big data analysis, types of big data analysis, benefits and barriers to big data analysis.*

Keywords: *Big data analysis, Internet of Things, OLAP, Multidimensional.*

1. INTRODUCTION:

Big data is the term used for the data sets that are so large that traditional processing methods are inadequate and inefficient. Big data usually includes data having sizes beyond the capability of commonly used software's to capture, manage and process. Big Data Analytics shows the challenges of data that are too large, too unstructured, and too fast moving to be managed by traditional and current available methods. From businesses and research institutions to governments, organizations routinely generate data of unpredictable scope and complexity. Getting meaningful information and competitive advantages from large amounts of data has become increasingly important to organizations worldwide. Trying to efficiently extract the meaningful results from such data sources quickly and easily is challenging task. Thus, analytics has become important to realize the full value of Big Data to improve their business performance. The tools available to handle the volume, velocity, and variety of big data have improved greatly in recent few years **[1]**. In general, these technologies are not much expensive, and much of the softwares are open source. Hadoop, the most commonly used takes incoming streams of data and distributes them onto cheap disks; it also provides tools for analyzing the data. However, these technologies require skill set, needed to work hard to collaborate all the important internal and external sources of data.

2. BIG DATA ANALYTICS:

Figure1: Big data analysis process **[8].**

The analysis of big data mainly involves analytical methods for big data, analytical architecture for big data, and analysis of big data. Data analysis is the final and the most important issue in the value chain of big data, with the purpose of extracting useful result values, providing suggestions or decisions. Different levels of potential values can be generated by the analysis of data values in various field. However, data analysis is a broad area, which changes overtime and is complex. In the following section, we introduce the types, benefits and barriers for big data analysis. **[3]**

Traditional data analysis

Traditional data analysis make use of proper statistical methods to analyze massive data, in order to concentrate, extract, and filter valuable data hidden in logs of large datasets, and to get the valuable data, so as to maximize the potential of data. Data analysis plays a vital role in making development plans for a country, helping in understanding customer demands for commerce, and analyzing market for enterprises. Big data analysis can be seen as the analysis technique for a special kind or foe large amount of data. Therefore, many traditional data analysis methods are still used for big data analysis. Some of the traditional data analysis methods are discussed in the following sections, many of which belong to statistics and computer science. **[5]**

***Cluster Analysis*:** It is a statistical method for grouping objects, and specifically, classifying objects according to some features. Cluster analysis is used to differentiate objects with particular features and divide them into some categories (clusters) according to these features, such that objects in the same category will have high homogeneity while different categories will have high heterogeneity.

***Factor Analysis*:** It is basically targeted at describing the relation among many elements with only a few factors, i.e., grouping several closely related variables into a factor, and the few factors are then used to reveal the most information of the original data.

***Correlation Analysis*:** It is an analytical method for determining the law of relations, such as correlation, correlative dependence, and mutual restriction. Such relations can be classified into two types: (i) Function, representing the strict dependence relationship among phenomena, which is also called a definitive dependence relationship; (ii) correlation, some undetermined or inexact dependence relations, and the numerical value of a variable may correspond to several numerical values of the other variable, which represent a regular fluctuation surrounding their mean values.

***Regression Analysis*:** It is a mathematical tool for revealing correlations between one variable and several other variables. Based on a group of experiments or observed data, regression analysis identifies dependence relationships among variables hidden randomly. Regression analysis may make complex and undetermined correlations among variables to be simple and regular.

***A/B Testing*:** It is also called bucket testing. It is a technology for determining how to improve target variables by comparing the tested group.

***Statistical Analysis*:** Statistical analysis, a branch of applied mathematics, randomness and uncertainty are modeled with Probability Theory. Statistical analysis can provide a description and an inference for big data. Descriptive statistical analysis can summarize and describe datasets, while inferential statistical analysis helps to draw conclusions from data subject to random variations. Statistical analysis is very much used in the economic and medical fields.

***Data Mining Algorithms*:** Data mining is a process for extracting hidden, unknown, but useful information and knowledge from massive, incomplete, noisy, fuzzy, and random data. In 2006, The IEEE International Conference on Data Mining Series (ICDM) identified ten most influential data mining algorithms through a strict selection procedure, which includeC4.5, k-means, SVM, Apriori, EM, Naive Bayes, and Cart, etc. These ten algorithms cover classification, clustering, regression, statistical learning, association analysis, and linking mining **[6]**.

Big data analytic methods

In the dawn of the big data era, people are concerned how to easily get key information from large data so as to bring values for enterprises and individuals. At present, some of the main processing methods of big data are shown as follows.

***Bloom Filter*:** Bloom Filter consists of a series of Hash functions. The principle of Bloom Filter is to store Hash values of data along with data itself by using a bit array, which is in essence a bitmap index that uses Hash functions to conduct lossy compression storage of data. It has advantages such as high space efficiency and high query speed, but also has some disadvantages in misrecognition and deletion.

***Hashing*:** It is a method that essentially transforms data into shorter fixed-length numerical values or index values. Hashing has such advantages as rapid reading, writing, and high query speed, but it is hard to find a good Hash function.

***Index*:** Index is always an effective method to reduce the cost of disk reading and writing, and improve insertion, deletion, modification, and query speeds in both traditional relational databases that manage structured data, and other technologies that manage semi structured and unstructured data.

***Triel*:** It is also called trie tree, a variant of Hash Tree. It is mainly applied to rapid retrieval and word frequency statistics. The main idea of Triel is to utilize common prefixes of character strings to reduce comparison on character strings to the greatest extent, so as to improve query efficiency.

***Parallel Computing*:** Parallel computing refers to simultaneously using many computation resources to complete a computation task. Its basic idea is to divide a problem and assign them to several separate processes to be independently completed, so as to achieve co processing. Presently, some of the parallel computing models are MPI (Message Passing Interface), Map Reduce, and Dryad which, are useful for big data analysis. Therefore, some high-level parallel programming tools or languages are being developed based on these systems. Such high-level languages include Sawzall, Pig, and Hive used for Map Reduce, as well as Scope and DryadLINQ used for Dryad. **[2]**

3. OBJECTIVES:

The main objectives of this paper are

- The research on Big Data technology is very mindful and brilliant.
- This makes us aware and enlightened our knowledge.
- Moreover, we researched about theory and this theory will definitely help in our practical implementation of Big Data Analytics.
- Theory is important in practical purposes.

4. TYPES OF DATA ANALYSIS:

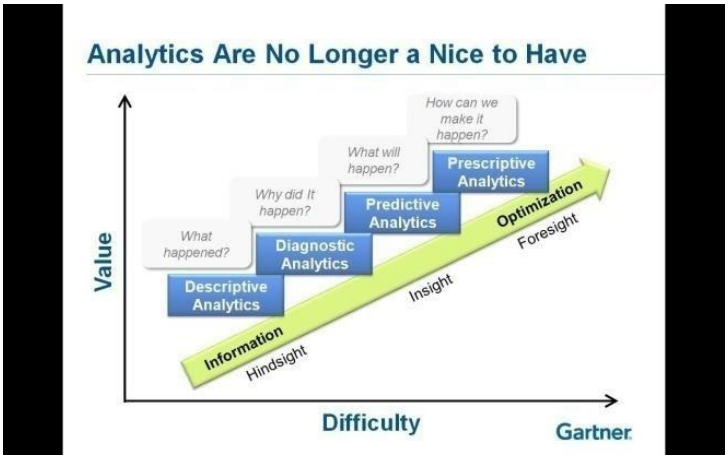

Figure 2.Analysis of Big Data **[9].**

Data analysis research can be divided into six technical fields, i.e., structured data analysis, text Data analysis, web data analysis, multimedia data analysis, network data analysis, and mobile data analysis. The main objective of this classification is to focus on data characteristics, but some of the fields may utilize similar basic technologies. Since data analysis has a broad scope and it is not easy to have a comprehensive coverage, so we will lay emphasis on the key problems and technologies in data analysis in the following discussions.**[6]**

Structured data analysis

Business applications and scientific research may generate massive structured data, of which the management and analysis rely on mature commercialized technologies, such as RDBMS, data warehouse, OLAP, and BPM (Business Process Management). Data analysis is mainly based on data mining and statistical analysis. However, data analysis is still a very active research field and new application demands the development of new methods. For example, statistical machine learning based on exact mathematical models and powerful algorithms have been applied to anomaly detection and energy control. Exploiting data characteristics, time and space mining can extract knowledge structures hidden in high-speed data flows and sensors. Driven by privacy protection in e-commerce, e-government, and health care applications, privacy protection data mining is an emerging research field.

Text data analysis

The most common format of information storage is text, e.g., emails, business documents, web pages, and social media. Therefore, text analysis emphasis on business-based potential than structured data. Generally, text analysis is a process to extract useful information and knowledge from unstructured text. Text mining is Inter-disciplinary, involving information retrieval, machine learning, statistics, computing linguistics, and data mining in particular. Most text mining systems are based on text expressions and natural language processing (NLP), with more emphasis on the latter. NLP allows computers to analyze, interpret, and even generate text.

Web data analysis

Web data analysis aims to automatically retrieve, extract, and evaluate information from Web documents and services so as to discover useful knowledge. Web analysis is related to several research fields, including database, information retrieval, NLP, and text mining. We can classify Web data analysis into three related fields: Web content mining, Web structure mining, and Web usage mining. Web content mining is the process to discover useful knowledge in Web pages, which generally involve several types of data, such as text, image, audio, video, code, metadata, and hyperlink. Since most Web content data is unstructured text data, the research on Web data analysis mainly centers on text and hypertext. Hypertext mining involves the mining of the semi-structured HTML files that contain hyperlinks. Supervised learning and classification play important roles in hyperlink mining, e.g., email, newsgroup management, and Web catalogue maintenance. Web structure mining involves models for discovering Web link structures. Here, the structure refers to the schematic diagrams linked in a website or among multiple websites. Models are built based on topological structures provided with hyperlinks with or without link description. Such models show the similarities and correlations between different websites and are used to classify them. Page Rank and CLEVER techniques make full use of the models to look up relevant website pages. Topic oriented crawler is a tool utilizing these models. Web usage mining aims to mine and analyze auxiliary data generated by Web activities. Web content mining and Web structure mining use the master Web data. Web usage data contains access logs at Web servers and proxy servers, browsers' history records, user profiles, registration data, user sessions or trades, cache, user queries, bookmark data, mouse clicks and scrolls, and any other kinds of data generated through interaction with the Web. **[4]**

Multimedia data analysis

Multimedia data (mainly including images, audio, and videos) have been growing at an amazing speed. Because multimedia data is heterogeneous in nature and have more key information than simple structured data or text data, extracting information from it faces the huge challenge of the semantic differences. Research on multimedia analysis includes fields such as multimedia summarization, multimedia annotation, multimedia index and retrieval, multimedia suggestion, and multimedia event detection, etc. Audio summarization can be accomplished by extracting the prominent words or phrases from metadata or synthesizing a new representation. Video summarization is to interpret the most important or representative video content sequence, and it can be static or dynamic. Static video summarization methods utilize a key frame sequence or context-sensitive key frames to represent a video. Such methods are simple and have been used in many business applications (e.g., by Yahoo, AltaVista and Google), but their performance is poor. Dynamic summarization methods make use of a series of video frame to represent a video, and take other smooth measures to make the final summarization look more natural. In, the authors propose a topic-oriented multimedia summarization system (TOMS) that can automatically summarize the important information in a video belonging to a certain topic area, based on a given set of extracted features from the video. Multimedia annotation inserts labels to describe contents of images and videos at both syntax and semantic levels. With such labels, the management, summarization, and retrieval of multimedia data can be implemented easily. Since manual annotation is time and labor intensive, automatic annotation without any human interventions becomes highly appealing. The main challenge for automatic multimedia annotation is the semantic difference. Multimedia indexing and retrieval can be define as describing, storing, and organizing multimedia information and assisting users to quickly look up multimedia resources. Generally, multimedia indexing and retrieval include five procedures: structural analysis, feature extraction, data mining, classification and annotation, query and retrieval. Structural analysis aims to segment a video into several semantic structural elements, including lens boundary detection, key frame extraction, and scene segmentation, etc. As per the result of structural analysis, the second procedure includes feature extraction, which contains mining, the features of key frames, objects, texts, and movements, which lay the foundation of video indexing and retrieval. Data mining, classification, and annotation are to utilize the extracted features to find the modes of video contents and put videos into scheduled categories so as to generate video indexes. On getting a query, the system will use a similarity measurement method to look up a candidate video. The retrieval result helps to optimize the related feedback. The content-based methods identify general features of users or their interesting, and recommend users for other contents with similar features. These methods largely rely on content similarity measurement, but most of them are challenged by analysis limitation and excess specifications. The collaborative-filtering-based methods identify groups with similar interests and recommend contents for group members according to their behavior. The research on video event detection is new, and mainly focuses on sports or news events, running or abnormal events in monitoring videos, and other similar events with repetitive patterns.

Network data analysis

Network data analysis evolved from the initial quantitative analysis and sociological network analysis into the emerging online social network analysis in the beginning of 21st century. Many online social networking services include Twitter, Face book, and LinkedIn, etc. have become popular over the years. Such online social network

services generally include massive linked data and content data. The linked data is in the form of graphic structures, which describe the communications between two entities. The content data have text, image, and other network multimedia data. The rich content in such networks brings about both unprecedented challenges and opportunities for data analysis. In consideration with the data-centered perspective, the existing research on social networking service contexts may be classified into two categories: link-based structural analysis and content-based analysis. The research on link-based structural analysis has always been done on link prediction, community discovery, social network evolution, and social influence analysis, etc. SNS can be seen as graphs, in which every vertex corresponds to a user and edges correspond to the correlations among users. Since SNS are dynamic networks in which new vertexes and edges are added continually to the graphs. Link prediction is used to predict the possibility of future connection between two vertexes. Some of the techniques used for link prediction, includes feature-based classification, probabilistic methods, and Linear Algebra. Feature-based classification is to select a group of features for a vertex and utilize the existing link information to generate binary classifiers to predict the future link. Probabilistic methods focus on to build models for connection probabilities among vertexes in SNS. Linear Algebra computes the similarity between two vertexes according to the singular similar matrix. A community can be represented by a sub graphic matrix, in which edges connected to vertexes in the sub-graph have high density while the edges between two sub-graphs feature much lower density many methods for community detection have been proposed and studied, most of which are topology-based target functions relying on the concept of capturing community structure. The objective of research on SNS aims to look for a law and deduction model to interpret network evolution. With the help of empirical studies it is found that proximity bias, geographical limitations, and other factors play important roles in SNS evolution, and some generation methods are proposed to assist network and system design. Social influence refers to the case when individuals change their behavior under the influence of others. The strength of social influence depends on the relation among individuals, network distances, time effect, and characteristics of networks and individuals, etc. Marketing, advertisement, recommendation, and other applications can benefit from social influence by qualitatively and quantitatively measuring the influence of individuals on others. Content-based analysis in SNS is also known as social media analysis. Social media include text, multimedia, positioning, and comments. However, social media analysis faces unpredicted challenges. Firstly, large and continually growing social media data should be automatically analyzed within a reasonable time span and Secondly social media data contains much noise. For example, blog sites have a large number of spam blogs, and so does trivial Tweets in Twitter. Third, SNS are dynamic networks by nature and are frequently varying and updated. The existing research on social media analysis is still new and finding its ways.

Mobile data analysis

By April 2013, Android Apps has provided more than 650,000 applications, covering about all categories. By the end of 2012, the monthly mobile data flow has reached 885 PB. The huge data and large number of applications needed mobile analysis, but also bring challenges along with it. As a whole, mobile data has unique characteristics, e.g., mobile sensing, moving flexibility, noise, and a large amount of redundancy. Recently, new research on mobile analysis has been started in various fields. Since the research on mobile analysis is in its infancy, we will only introduce some recent and representative analysis applications only. With the growth of more mobile users and improved performance, mobile phones are now helpful in building and maintaining communities which includes communities with geographical locations and communities based on different cultural backgrounds and interests (e.g., the latest Web chat). Traditional network communities or SNS communities are in few numbers with online interactions among members, and these communities are active only when members are operating before computers. On the contrary, mobile phones can support rich interaction at anytime and anywhere. Mobile communities are defined as that a group of individuals with the same hobbies (i.e., health, safety, and entertainment, etc.) gather together on networks, meet to make a common goal, decide measures through consultation to achieve the goal, and start to implement their plan. It is now believed that mobile community applications will promote the development of the mobile industry. Recently, with the progress in wireless sensor, mobile communication technology, and stream processing help people to build a body area network to have real-time monitoring of people's health. Generally, medical data among various sensors have different characteristics in terms of attributes, time and space relations, along with physiological relations, etc and also such datasets involve privacy and safety protection. In, Garg et al. introduce a multi-modal transport analysis mechanism of raw data for real-time monitoring of health. Under the circumstance that only highly comprehensive characteristics related to health are available, Park et al. in examined approaches to better utilize. Researchers from Gjovik University College in Norway and Derawi Biometrics jointly developed an application for smart phones, analyzing paces when people walk and uses the pace information for unlocking the safety system.

5. WHY PUT BIG DATA AND ANALYTICS TOGETHER NOW?

Big data provides large statistical cases, which improve analytic tool results. Most of the tools constructed for data mining or statistical analysis are used for large data sets. In fact, the familiar rule is that the larger the data sample, the more accurate are the statistics and other products of the analysis. Instead of using mining and statistical tools, many users generate hand-code complex SQL, which scan big data in searching right customer segment. The new generation of data visualization tools and in-database analytic functions are used to operate on big data. Analytic tools can also execute big queries and scan tables in record time. Recent generations of local tools and platforms have lifted us onto a new plateau of performance that is very compelling for applications involving big data. The economics of analytics is now more uplifted than ever. This is due to a drop in the cost of data storage and processing bandwidth. The fact that tools and platforms for big data analytics are relatively affordable and is significant because big data is not just for only big business. Many small-to-midsize businesses also need to manage and leverage big data. Most of the modern tools and techniques used for advanced analytics and big data are very tolerant of raw source data, with its transactional schema, non-standard data, and poor-quality data because discovery and predictive analytics depend on lots of detail. For example, analytic applications for fraud detection often depend on outliers and deviant data as indications of fraud. So, be careful: If you apply ETL and data quality processes to big data as you do for a data warehouse, you run the risk of stripping out the very nuggets that make big data a gem for advanced analytics. Big data is a special asset that merits leverage. That's the real point of big data analytics. The new technologies and new best practices are fascinating, even mesmerizing, and there's a certain macho coolness to working with dozens of terabytes. But don't do it for the technology. Analytics based on large data cases shows business change. The recession has lead to increase the pace of business. The recovery brings even more change. The average business has changed beyond all acceptances because of the new economic recession and recovery. The change has not gone unnoticed. Business people now share a wholesale recognition that they must explore change just to understand the new state of the business.

6. BENEFITS OF BIG DATA ANALYTICS:

We saw that user organizations have adopted big data analytics in appreciable numbers. To determine the potential benefits that are driving the adoption, TDWI's survey asked: "Which of the following benefits would ensue if your organization implemented some form of big data analytics?" The most of the benefits are those which are often selected by survey respondents and the benefit declines while list moves downward anything involving customers could benefit from big data analytics. At the top of the list, this includes better-targeted social-influencer marketing, customer-base segmentation, and recognition of sales and market opportunities. Late economic changes worldwide have changed consumer perception. Big data analytics help in developing definitions of churn and other customer behaviors, along with an understanding of consumer behavior from click streams. Business intelligence can benefited from big data analytics which could result in more number of and accurate business insights, an understanding of business change, better planning and forecasting, and the identification of root causes of cost. Specific analytic applications are likely beneficiaries of big data analytics. For example, consider analytic applications for the detection of fraud, the quantification of risks, or market sentiment trending. At the leading edge, big data analytics can help automate decisions for real-time business processes such as loan approvals or fraud detection. Potential benefits entered by survey respondents selecting "other" include customer loyalty, service experience optimization, healthcare delivery optimization, and supplier performance based on cost and quality. **[7]**

7. BARRIERS TO BIG DATA ANALYTICS:

To have a sense of which barriers are more likely than others, this report's survey asked: "In your organization, what are the top barriers hindering the implementing of big data analytics?" The most likely barriers are discussed as follow. Problems with skills, sponsors, and software are the leading barriers. Inadequate staffing are the leading barriers to big data analytics. As, many organizations are still new to big data analytics and their skill set is not quite the same as that needed for business intelligence and data warehousing, for which most organizations have developed their own skills. Other barriers have the difficulty of constructing a big data analytic system and problems with making big data usable for end users. A lack of business support can halt a big data analytics program. Lack of business sponsorship and a lack of a compelling business case, with the related issue of overall cost can also be the barriers. Problems with database software can be barriers to big data analytics. Problem arise when the current database software doesn't have in-database analytics, have scalability problems with big data, can't process analytic queries fast enough, or cannot load data fast enough. In a related issue, managing big data in a data warehouse is challenging when that warehouse is meant for reports and OLAP only. Other Possible barriers may include competing with other initiatives, lack of test and control rigor, and sourcing. **[7]**

8. CONCLUSION:

This paper is basically reviewed about the different aspects of big data analysis. What is big data analysis and how it can be done. It deals with big data analytic methods including traditional and currently available ones, types of big data analysis, benefits and barriers to big data analysis. Yet it is not over, there are still new future aspects and

challenges of big data are discovered day by day to improve analysis and data mining to get best fit tools to handle it and get maximum results out of it.

REFERENCES:

1. Fan,W & Bifet,A . Mining big data: Current status and forecast to the future. *SIGKDD Explorations*, 14(2), 1-5.
2. Snijders,C, Matzat,U & Reips,U. Big Data: Big Gaps of Knowledge in The Field of Internet.*International Journal of Internet Science*, 7, 1-5.
3. Cuzzocrea,A, Song,Y & Davis,K.(2011). Analytics of Large Scale Multidimensional Data: The Big Data Revolution!.*DOLAP*.11.
4. Aggarwal,C.(2013).Managing and Mining Sensor Data. *International Journal of Advances In Database System, Springer*.
5. Bifet,A, Holmes,G, Kirkkby,R, & Pfahringer,B (2010).MOA: Massive Online Analysis. *International Journal of Machine Learning Research.*
6. Chen.M, Mao.S, & Liu.Y(2014). *Big Data: A Survey*. *International Journal of science and business media, Springer*.
7. Russom.P (2011). Big data Analytics.*TDWI Best practices Report, TDWI Research*
8. datanami.(n.d.).Google,Big data analysis process. [Infographic]. Retrieved from https://www.datanami.com/2013/10/07/six_steps_to_extract_value_from_big_data/
9. blogs.gartner (n.d.).Analysis of big data. [Infographic].Retrieved from https://blogs.gartner.com/matthew- davis/top-10-moments-from-gartners-supply-chain-executive-conference/