DATA ANALYSIS: METHODS, TOOLS, AND CHALLENGES
Submitted by: Pranavi Ganti

ABSTRACT

In today's digital age, enormous amounts of data have become available on hand to decision-makers. Big Data refers to massive data sets that may be analyzed computationally to reveal patterns, trends, and associations. Various fields like healthcare, banking, finance, and cyber technologies generate terabytes of data every day. The extraction of data for decision-making from these enormous data sets necessitates a lot of effort at various levels. As a result, big data analysis is an emerging area for research and development. The primary goal of this paper is to study multiple stages of data analytics, the challenges faced, open research questions and the tools used. As a result, this article serves as a starting point for investigating big data at various phases. Furthermore, it gives up a new avenue for academics to develop a solution to overcome the obstacles

CHAPTER 1

1.1 INTRODUCTION

Data is generated from numerous sources in the digital age, and the rapid transition from digital technologies has resulted in the creation of big data. With a huge dataset collection, it permits evolutionary breakthroughs in a variety of fields. Essentially, it refers to a collection of huge and complicated datasets that are difficult to analyze using typical database administration tools or data processing programs.[1] These are available in petabytes and beyond in structured, semi-structured, and unstructured formats.[2] There are two types of data sources: primary and secondary. Primary data is assembling raw data or information. Whereas, secondary data, which are normally received from many departments of the country such as health, education, and population and may be collected from various hospitals, clinics, and schools' records, might be used for our own research.

The use of secondary data may save our survey costs and time, and it may be reliable if the information was obtained by a government body. However, it has a number of drawbacks. The secondary data may be out of date for the analysis we wish to do. On the other hand, the primary source of data has greater

---

[1] Muhammad Ibrahim, *The art of Data Analysis*, (2015).
[2] Nada Elgendy & Ahmed Elragal, *Big Data Analytics: A Literature Review Paper*, (2014).

advantages than the secondary source. Surveys, focus groups, questionnaires, personal interviews, experiments, and observational studies can all be used to collect data which might be more accurate comparatively. The primary goal of big data analysis is to handle large amounts of data utilizing a variety of traditional and computational intelligent techniques. [3]The processing of data and further analysis is broken up into three stages:

- Data management
- Exploratory data analysis
- Statistical Data analysis

Data warehouses are commonly used to manage huge datasets. The extraction of exact   knowledge from available huge data is the most pressing challenge in this scenario. Most of the methodologies given in data mining are incapable of successfully handling huge datasets. [4]The main issue with big data analysis is a lack of coordination across database systems as well as analysis techniques such as data mining and statistical analysis. These difficulties typically arise when we want to do knowledge discovery and presentation for practical applications. However, it should be highlighted that not all data provided in the form of big data is relevant for analysis or decision making.

1.2 RESEARCH OBJECTIVES

- Types of data for Analysis
- Methods used for Data Analysis
- Process and tools for Data Analysis
- Big Data challenges

1.3 RESEARCH QUESTIONS

This paper answers the following questions:

1. What is big data?
2. What are the types of data?
3. What is data analysis and what are the methods and techniques used for data analysis?
4. What is the process of data analysis and what are the tools used?
5. Is big data really helpful for data analysis?

---

[3] Wesely Chai, *Big Data Analytics,* (2016).
[4] Debi Prasanna Acharjya & Kauser Ahmed P, *A Survey on Big Data Analytics: Challenges, Open Research, Issues and Tools*, International Journal of Advanced Computer Science and Applications, (2016).

6. What are the challenges for big data analysis?

CHAPTER 2

2.1 TYPES OF DATA FOR ANALYSIS

Every type of data has the unique property of describing objects after being assigned a specific value. To make analysis helpful, these values must be organized, processed, and presented in a specific context. Data can take many forms, here are the most common data types.

- Qualitative Data
- Quantitative Data

2.1.1 Qualitative Data

When the data supplied includes words and descriptions, we refer to it as qualitative data[5]. Although this data can be observed, it is subjective and difficult to examine in study, particularly for comparison. Everything describing flavor, sensation, texture, or an opinion that is deemed quality data is represented by quality data. This type of information is typically gathered through focus groups, personal interviews, or open-ended questions in surveys. Because qualitative data is made up of words, descriptions, images, objects, and sometimes symbols, data analysis and qualitative data research function a little differently than numerical data.[6] Obtaining understanding from such complex material is a difficult process. As a result, it is commonly utilized for exploratory research and data analysis. Qualitative research is thought to be extremely useful for acquiring a thorough knowledge of underlying reasons and motivations.[7] It gives information about the context of a problem. Simultaneously, it frequently creates ideas and hypotheses for subsequent quantitative study. It is critical in determining the frequency of specific features or qualities. It enables statisticians or researchers to create parameters that can be used to examine larger data sets.

The use of qualitative data allows observers to quantify the environment around them. In this research, data analysis is defined as the process of systematically finding and organizing interview transcripts, observation notes, or other non-textual resources accumulated by the researcher in order to gain a better understanding of the phenomenon. The majority of the time, the process of analyzing qualitative data

---

[5] https://www.questionpro.com/blog/qualitative-data/
[6] https://www.questionpro.com/blog/data-analysis-in-research/
[7] https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/datopic.html

entails coding or categorizing the data. Essentially, it entails making sense of massive volumes of data by lowering the volume of raw information, then discovering relevant patterns, pulling meaning from data, and ultimately establishing a logical chain of evidence. [8]The most crucial stage in the qualitative data analysis process is coding or categorizing the data. Coding and data analysis are not synonymous, despite the fact that coding is an important part of the qualitative data analysis process. Coding is just the process of subdividing a massive amount of raw information or data and categorizing it. In non-scientific terms, codes are tags or labels used to assign discovered themes or subjects from the study's data. Traditionally, coding was done by hand, using coloured pens to categorize data and then cutting and sorting the data. With the improvement of software technology, qualitative researchers are increasingly using electronic data coding methodologies. Therefore, the use of computers in qualitative analysis merely made organization, reduction and storage of data more efficient and manageable.

2.1.2 Quantitative Data

Quantitative data is any data that is expressed as a series of numerical figures. This form of data can be classified, grouped, measured, calculated, or rated. [9]This form of data includes questions such as age, rank, cost, length, weight, scores, and so on. You can present such data in graphical form, such as charts, or you can use statistical analytic tools to analyze it. Because of the ease of mathematical derivation that quantitative data provides, it makes measuring various parameters controllable. Quantitative data is typically collected for statistical analysis through the use of surveys, polls, or questionnaires distributed to a specified segment of a population.[10] The obtained results can be applied to a population.

Quantitative Data can be divided into two types:

- Administrative Data: Basic monitoring techniques generate administrative data. It is frequently concerned with activities or outputs, such as the number of training sessions held or children immunized. It may also include financial or logistical considerations.[11]

- Sampled or Population Data: Sampled or population data, on the other hand, is frequently collected to examine changes brought about by a project or programme. It typically includes data from a sample (of people, homes, towns, events, etc.) or an entire population, like as all of the farmers in a region.

---

[8] https://www.questionpro.com/blog/qualitative-data/
[9] https://www.questionpro.com/blog/data-analysis-in-research/
[10] https://www.questionpro.com/blog/quantitative-data/#Advantages_of_Quantitative_Data
[11] https://www.intrac.org/wpcms/wp-content/uploads/2017/01/Quantitative-analysis.pdf

Many stakeholders prefer quantitative data to qualitative data when making decisions. This is due to a number of factors.[12] To begin, quantitative analysis norms are well recognized and well established. If these guidelines are followed correctly, a quantitative analysis should produce the same findings regardless of who performed the task. In contrast, qualitative analysis places a great deal of reliance on the talents and integrity of the analyst.

Second, the fact that quantitative studies can be duplicated eliminates the possibility of intentional bias. Anyone with access to the same data might theoretically produce the same results. This means that work can be double-checked and validated. This makes it far more difficult to modify conclusions to suit the individual or organization conducting the analysis. As the results obtained are objective in nature, they are extremely accurate.

2.2 METHODS FOR DATA ANALYSIS

Data analysis is the systematic application of statistical and logical techniques to describe the data scope, modularize the data structure, condense the data representation, illustrate through images, tables, and graphs, and evaluate statistical inclinations, probability data, and derive meaningful conclusions.[13] These analytical approaches allow us to derive the underlying inference from data by removing the extraneous confusion caused by its remainder. Data generation is a continuous activity, which makes data analysis a continuous, iterative process in which data collection and analysis occur concurrently. One of the most important aspects of data analysis is ensuring data integrity.

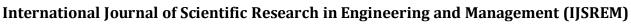The following are the main data analysis methods:

- Qualitative Analysis: This method primarily addresses issues such as 'why,' 'what,' and 'how.' Each of these issues is handled using quantitative techniques such as questionnaires, attitude scaling, standard outcomes, and others.[14] This type of analysis is typically presented in the form of texts and narratives, but it may also include audio and video representations.

- Quantitative Analysis: This study is generally measured in terms of numbers. The data presented here are in the form of measurement scales and can be further statistically manipulated.[15]

---

[12] https://www.questionpro.com/blog/quantitative-data/#Advantages_of_Quantitative_Data
[13] Surya R Niraula, *A review of Research process, data collection and analysis*, (2019).
[14] https://www.questionpro.com/blog/qualitative-data/
[15] https://www.intrac.org/wpcms/wp-content/uploads/2017/01/Quantitative-analysis.pdf

- Text Analysis: Text analysis is a method of analyzing text in order to extract machine-readable data. Its goal is to generate organized data from unstructured and free resources. The procedure entails slicing and dicing massive amounts of unstructured, heterogeneous data into easy-to-read, handle, and analyze data chunks.[16] Text mining, text analytics, and information extraction are other terms for it.

- Statistical Analysis: Statistic methods is the gathering, analysis, and validation of data. It is the process of executing many statistical operations on data in order to quantify it and apply statistical analysis.[17] Descriptive data, such as surveys and observational data, are examples of quantitative data. It's also known as a descriptive analysis. It includes various tools to perform statistical data analysis such as SAS (Statistical Analysis System), SPSS (Statistical Package for the Social Sciences), Stat soft, and more.

- Diagnostic Analysis: Diagnostic analysis is a step beyond statistical analysis in order to provide more in-depth analysis to answer the queries. It is also known as root cause analysis since it encompasses techniques such as data discovery, mining, drill down, and drill through.

- Predictive Analysis: Predictive analysis analyses historical data loaded into a machine learning model to identify significant patterns and trends. [18]To anticipate what happens next, the model is applied to the current data. Many companies prefer it because of its numerous advantages, such as data volume and kind, quicker and cheaper computers, user-friendly software, tighter economic constraints, and a desire for competitive distinctiveness.

- Prescriptive Analysis: Prescriptive analytics recommends several courses of action and illustrates the potential outcomes that could result from predictive analysis. Prescriptive analysis, which generates automated conclusions or suggestions, necessitates a specific and distinct algorithmic approach, as well as explicit instruction from people who use the analytical tools.

2.3 PROCESS AND TECHNIQUES FOR DATA ANALYSIS

In a nutshell, data analysis is a process of putting facts and numbers together to answer a search problem. It is critical to obtaining answers to the research question. Another important aspect of the research is the interpretation of the data, which is derived from the data analysis and draws inferences and conclusions.

---

[16] https://hackr.io/blog/what-is-data-analysis-methods-techniques-tools
[17] https://www.academia.edu/8135057/Methods_of_Data_Analysis
[18] https://www.datapine.com/blog/data-analysis-methods-and-techniques/

The data analysis process involves acquiring all relevant information, processing it, studying the data, and applying it to discover patterns and other insights.[19]

The process consist of the following phases:

- Data requirement Analysis: The data requirements analysis process uses a top-down strategy that prioritizes business-driven demands. Therefore the analysis is carried out to guarantee that the identified requirements are relevant and practical. The process includes data search and evaluation in the context of expressly qualified business data consumer requirements.[20]

- Data Collection: Data collection is the process of acquiring and measuring information on variables of interest in a systematic manner in order to answer specified research questions, test hypotheses, and evaluate outcomes[21]. Data gathering is a component of research that is shared by all fields of study, including physical and social sciences, humanities, business, and so on.

- Data Cleaning: Data cleaning is the process of removing data from your dataset that does not belong there. This takes place along with Data Transformation. The process of changing data from one format or structure to another is known as data transformation. Transformation processes, often known as data wrangling or data munging, involve changing and mapping data from one "raw" data type to another for warehousing and analysis.

- Data Analysis/ Processing: In this phase, data analysis and other software tools are used to interpret and analyze data to arrive at conclusions[22]. Data Analysis tools such as Excel, Python, PowerBI, R, Matlab are used to carry forward this process.

- Data Interpretation: Data interpretation is the process of reviewing data using specified processes in order to give meaning to the data and reach a suitable conclusion. It involves taking the results of data analysis, drawing conclusions from the relationships studied, and applying them to reach a conclusion. This further contains two processes, namely

- ➢ Data Mining
- ➢ Data Modeling

- Data Visualization: The graphical depiction of information and data is known as data visualization[23]. Data visualization tools, which include visual components such as charts, graphs, and maps, give an easy way to

[19] https://hackr.io/blog/what-is-data-analysis-methods-techniques-tools
[20] https://www.datapine.com/blog/data-analysis-methods-and-techniques/
[21] Joel Ashirwadam J. W, *Methods of Data Analysis*, (2017).
[22] https://hackr.io/blog/what-is-data-analysis-smethods-techniques-tools
[23] https://hackr.io/blog/what-is-data-analysis-methods-techniques-tools

observe and comprehend trends, outliers, and patterns in data. Data visualization tools and technologies are essential in the era of big data for analysing enormous volumes of data and making data-driven decisions.

2.3.1 Data Analysis Tools:

There are numerous data analysis tools, each with its own set of features. The tools used should always be chosen based on the sort of analysis and data being dealt with. Here is a list of some interesting Data Analysis tools.

- Excel: It provides a wide range of appealing features, and with additional plugins installed, it can handle large amounts of data. So, if your data isn't close to the substantial data margin, Excel can be a useful tool for data analysis.

- Tableau: It is classified as a BI Tool as it is developed solely for data analysis. Tableau's essence is the Pivot Table and Pivot Chart, which work to show data in the most user-friendly way possible. It also offers a data cleansing feature as well as analytical skills functions.

- Power BI: It began as an Excel plugin before branching out to become one of the most powerful data analytics tools. It is available in three different versions: free, pro, and premium. Its PowerPivot and DAX languages can do sophisticated advanced analytics in the same way that Excel formulas can.

- R and Python: These are extremely powerful and adaptable programming languages. R excels at statistical analysis tasks like the normal distribution, cluster classification techniques, and regression analysis.[24] Individual predictive studies are also performed, such as consumer behavior, expenditure, things favored by him based on his browsing history, and more.

- SAS: It is a data analytics and data manipulation programming language that can readily retrieve data from any source. SAS has developed a wide range of customer profiling products for web, social media, and marketing analytics. It can anticipate their actions, manage, and optimize communications.

2.4 BIG DATA CHALLENGES

Data volumes are growing, as are the possibilities for what may be done with so much raw data. [25]While Big Data has numerous advantages, it also has a number of drawbacks. This is a new set of complicated

---

[24] https://www.proschoolonline.com/blog/top-10-data-analytics-tools
[25] https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

technologies that is still in its early stages of development and evolution. The top critical challenges of Big Data are as follows:

- Data Storage and Analysis: The size of data has expanded tremendously in recent years due to numerous ways such as mobile devices, aerial sensory technologies, remote sensing, radio frequency identification readers, and so on.[26] These data are stored at a high expense, yet they are eventually ignored or deleted due to a lack of storage space. As a result, the first issue for big data analysis is the availability of storage mediums and faster input/output speeds. In such instances, data accessibility must be prioritized for knowledge discovery and representation.

- Complications in knowledge and computation: The discovery and representation of knowledge is a critical issue with big data. It is divided into several subfields, including authentication, archiving, management, preservation, information retrieval, and representation.[27] As the size of big data is growing at an exponential rate, the present technologies may not be efficient in processing it to derive relevant information. Data warehouses and data marts are the most prevalent approaches for managing massive datasets. A data warehouse is primarily responsible for storing data generated from operational systems, whereas a data mart is built on a data warehouse[28]. However, present large data analysis methods perform poorly when it comes to dealing with computational complexities, uncertainty, and inconsistency. It is a significant challenge to design strategies and systems that can effectively deal with computational complexity, ambiguity, and inconsistencies.

- Data Scalability and Visualization: The biggest challenge for large data analysis techniques is scalability and security. The goal of data visualization is to convey data more effectively by applying graph theory techniques. Graphical depiction connects data with appropriate interpretation. However, most existing big data visualization tools perform poorly in terms of functionality and scalability.

- Information Security: Massive amounts of data are correlated, examined, and mined for relevant patterns in big data analysis. Every organization has distinct rules in place to protect its sensitive information. The preservation of sensitive data is a crucial challenge in big data analysis. Big data carries a significant security risk. As a result, information security is becoming a major data analytics issue. The most difficult task is to create multi-level security, a privacy-preserving data model for big data.

---

[26] Debi Prasanna Acharjya & Kauser Ahmed P, *A Survey on Big Data Analytics: Challenges, Open Research, Issues and Tools*, International Journal of Advanced Computer Science and Applications, (2016).
[27] https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0206-3
[28] K Venkatesh & S. Pasupathy, *Big Data Analytics: Challenges, Issues, Applications and Research Opportunity*, (2018).

CHAPTER 3

3.1 ANALYSIS AND CONCLUSION

Data is being generated at an alarming rate in recent years. To understand data, many data analysis methods and techniques are applied. Data analytics is likely to fundamentally alter how we live and conduct business in the future. We already employ analytics in our technology devices for many decisions in our daily lives. Analyzing this accurately is challenging for an ordinary individual. From this research, we studied the big data and data analytics processes, as well as their significance in decision making. As a result, big data and its significance were debated. Furthermore, various data analytics tools and approaches, in particular, were investigated. Big data methods and challenges were detailed. Using Data Analytics, important information may be retrieved and used to improve decision making and support informed decisions. Finally, when implemented appropriately, any new technology can bring with it a plethora of potential benefits and advances, let alone big data, which is a remarkable field with a promising future if treated correctly. However, big data is extremely difficult to manage. It necessitates proper storage, administration, integration, federation, cleansing, processing, and analysis, among other things. As a result, future research can focus on developing a roadmap or framework for large data management that addresses the existing challenges. Big data analytics is seen to be of tremendous importance in this era of data overflow and can give fresh insights and benefits to decision-makers in various fields. Big data analytics, when used correctly, has the ability to provide a foundation for scientific, technical, and humanitarian improvements.

BIBLIOGRAPHY

➢ Ibrahim, Muhammad. (2015), *The art of Data Analysis.*

➢ Ashirwadam J. W, Joel. (2017), *Methods of Data Analysis*.

➢ Acharya, Debi Prasanna. & Ahmed P, Kauser. (2016), *A Survey on Big Data Analytics: Challenges, Open Research, Issues and Tools.*

➢ https://www.academia.edu/8135057/Methods_of_Data_Analysis

➢ https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

➢ https://www.datapine.com/blog/data-analysis-methods-and-techniques/