Big Data Analytics: A Survey

Wasnaa Kadhim Jawad
Informatics Institute for Postgraduate Studies,
Iraqi Commission for Computers and Informatics,
Baghdad, Iraq
phd202120678@iips.icci.edu.iq

Abbas M. Al-Bakry
University of Information Technology and Communications
Baghdad, Iraq
abbasm.albakry@uoitc.edu.iq

Abstract

Internet-based programs and communication techniques have become widely used and respected in the IT industry recently. A persistent source of "big data," or data that is enormous in volume, diverse in type, and has a complicated multidimensional structure, is internet applications and communications. Today, several measures are routinely performed with no assurance that any of them will be helpful in understanding the phenomenon of interest in an era of automatic, large-scale data collection. Online transactions that involve buying, selling, or even investing are all examples of e-commerce. As a result, they generate data that has a complex structure and a high dimension. The usual data storage techniques cannot handle those enormous volumes of data. There is a lot of work being done to find ways to minimize the dimensionality of big data in order to provide analytics reports that are even more accurate and data visualizations that are more interesting. As a result, the purpose of this survey study is to give an overview of big data analytics along with related problems and issues that go beyond technology.

Index Terms: Big Data, Big Data Reduction, Big Data Analytics, Big Data Big Applications, Big Data Technologies

I. INTRODUCTION

The amount of data being generated exponentially increases as more and more gadgets and users access the internet. The vast amounts of information discovered on the Internet are measured in exabytes (EB) and petabytes (PB). Forecasts indicate that by 2025, the Internet's total capacity will have surpassed all human intelligence put together. The expansion of digital sensors, calculations, connections, and storage, which has led to large data aggregations, is primarily responsible for the proliferation of data. To explain this oddity, a researcher by the name of Roger Magoulas created the term "big data" [1].

The research company Gartner predicts that "information or data" will surpass oil in value in the twenty-first century. Data has exploded in popularity over the past 25 years, covering a wide range of fields and file types. The International Data Corporation (IDC) estimates that the amount of data created globally peaked in 2011 at 1.8 ZB and will expand by almost nine times over the next five years [2]. It is increasingly clear how important big data is in a variety of industries, including marketing, disease prevention, smart cities, and business intelligence [3].

Massive datasets are frequently characterized using the terminology of "big data" technology and related analytical approaches in light of the ever-increasing global data volume. Big data's semi- and unstructured types demand more frequent and in-depth real-time analysis as compared to standard datasets and their attendant procedures. A deeper understanding of the values that lie beneath the surface is made possible, and new challenges are offered, such as the necessity to uniquely organize and manipulate such massive datasets. As more and more data is made available from various sources, it also brings to light complicated issues that demand quick fixes. Data visualization is a crucial yet frequently disregarded component of big data analytics issues. This is a disadvantage because it will only be used to visualize the final report of data analytics [4].

The recent rapid development of information technology (IT) has made data generation easier. For instance, every minute, people post to YouTube enough new content to last 72 hours. Due to this data influx, businesses are being put to the test by the difficulties of getting and integrating vast amounts of data from scattered sources, including social networking apps.

The Internet of Things (IoT) and cloud storage have had phenomenal success, which is assisting with the tremendous growth of digital information. The cloud is the de facto standard for business preservation and retrieval of enormous data assets. The Internet of Things uses sensors to gather and transmit data for cloud archiving and analysis. The majority of businesses' present IT infrastructures, as well as the real-time processing and analytics that these systems are capable of, cannot handle these data sets because they are too large and complex. As a result, storing and retrieving enormous heterogeneous datasets has proven to be quite difficult and requires specialized hardware and software.

In order to give a high-level overview of big data analytics, this survey's goal is to do so. The extra organization of this literature study is summarized as follows: Section 2 provides an overview of big data analytics ideas and applications. The technologies that enable these applications are covered in

Section 3. In Section 4, issues that arose during the study are highlighted, along with solutions. Section 5 illustrates big data algorithms and is followed by a conclusion.

II. . RELATED WORKS

A five-node Hadoop cluster was created by Qin Yao to implement the distributed MapReduce algorithms. In comparison to individual nodes, our distributed algorithms demonstrate the potential to support effective data processing with huge amounts of medical data in healthcare services and clinical research. Additionally, by offering individualized recommendations, medical big data analytics enables us to develop hospital information systems that are smarter besides simpler to use [5].

M. MAZHAR touched on the problems of IoT with healthcare systems besides the problem of high volume medical data. M. MAZHAR suggested a Hadoop-based Intelligent Care System (HICS) that shows collaborative, contextual big data sharing based on IoT between all devices in the healthcare system. The suggested system consists of connected healthcare devices with a centre to gather massive data from these devices and analyze them using Hadoop techniques [6].

In order to address the issue of keeping several little files on the Hadoop bus, Hui He suggests a method. Typically, it combines small files and stores the resulting huge file. In order to efficiently minimize the blocks of HDFS data and lower the memory overhead of the master nodes in the cluster, Hui He devised a technique for merging tiny files depend on data block equilibrium. This method will enhance the size distribution of large files after merging [7].

To process large data in healthcare systems, numerous techniques and tools have been grown. The significance of big data in healthcare and the different tools existing in the Hadoop ecosystem to cope with it were discussed by Sunil Kumar. We also look at the theoretical framework of big data analytics in healthcare that comprises text/image, clinical decision support system, and electronic health records, besides the history of data collecting for various branches [8].

So as to speed up data processing while using up less storage space, Daoqu Geng presented an optimized platform for industrial big data depend on already existing big data frameworks. In particular, it concentrated on assessing how different compression and serialization techniques affected the functionality of the big data platform besides attempting to choose the best techniques for the industrial data platform. The trial results disclosed that the platform's data compression time was decreased by 73.9% with less than 96% of the compressed data volume and that the platform's data serialization time was decreased by 80.8% when compared to the combined methods of Hadoop and Spark. It takes less time to compare with measuring approaches as the amount of data increases [9].

As a result of the recent increase in the amount of data that exists in the healthcare industry, it now takes exceptional analytical skills to examine the data that is gathered from healthcare systems. Sabyasachi Dash created techniques for managing and offering pertinent answers to enhance public health. Healthcare professionals ought to have all the necessary infrastructure in place to generate and process large data in an organized manner. By creating new opportunities for modern healthcare, the appropriate administration, and analysis, besides the interpretation of big data, can completely alter the game [10].

Aishwarya Jaiswal compares different large data analysis technologies in his work, and he goes into great detail on Hadoop and Spark. Give an impression of the difficulties with big data, Spark, and Hadoop as well. Additionally, methods for resolving spark and Hadoop problems are extensively covered [11].

Explored Yasmine Benlachmi Since MapReduce has several flaws; Spark was developed for quick searches and real-time data streaming. The DAG and RDD techniques make up the spark rule. The major aim of this study is to assess the performance of Hadoop and Spark while comparing their fundamental features [12].

Mudasir Khan modified the Hadoop framework and deep learning classifier to develop a method of sentiment analysis. For feature extraction, a distributed data cluster using Hadoop is used. Next, utilizing data from Twitter, the crucial traits are retrieved. Each piece of input Twitter data is given a real-value review using a deep learning classifier, also known as a deep recurrent neural network classifier, which divides the input data into two categories, such as positive review and negative review. Metrics like rating accuracy and specificity are used to assess performance [13].

Lidong Wang introduced the main approaches, platforms, and tools for big data analytics in medical and healthcare engineering. Advancements and technological advances in big data analytics in healthcare are presented, that include artificial intelligence (AI) with big data, advanced computing and data processing, privacy, and cyber security, health economic outcomes, smart healthcare with sensors, and the Internet of Things. (IoT). Existing challenges of handling big data and big data analytics in healthcare are also presented in addition to future work.

The key techniques, frameworks, and resources for big data analytics in the fields of medical and healthcare engineering were introduced by Lidong Wang. Big data analytics for healthcare are advanced and technologically advanced, including artificial intelligence (AI) with big data, advanced computing and data processing, privacy, and cybersecurity, health economic outcomes, smart healthcare with sensors, and advanced computing and data processing (IoT). Future efforts, as well as current difficulties with big data analytics in the healthcare field, are presented [14].

TABLE I
SUMMARY OF RELATED WORK

Reference of Article	Years	Technology	Result
[5]	2015	Apache Hadoop MapReduce and Hadoop-based medical big data processing system	The system designed solves problems of MBD collection, storage, and analysis.
[6]	2017	Platform Hadoop and file merging and storage	After testing, the proposed technique showed that it increases the efficiency of data processing for groups while reducing memory consumption in their main nodes.
[7]	2017	Hadoop with big data in IoT Environment	Process high-speed WBAN sensory data in real time.
[8]	2018	Hadoop systems	Using a conceptual architecture with Hadoop-based terminology to solve healthcare problems in big data
[9]	2019	Industrial data analysis platform	Single-point failure is avoided by using a high-availability file storage system.
[10]	2019	Management and analysis of big data	Provided a set of solutions for users of healthcare systems, managing and analyzing data.
[11]	2020	Hadoop and Spark	An idea of how to overcome problems and challenges that occur in the Hadoop and Spark frameworks
[12]	2020	Hadoop and Spark	Explores a novel Apache Spark model technique as a replacement for the Hadoop MapReduce framework for efficient analysis of large amounts of HDFS data.
[13]	2020	Hadoop frameworks and deep learning classifier	The deep RNN approach built on Hadoop offered the highest levels of accuracy, sensitivity, and specificity.
[14]	2020	Big data analytics for healthcare include AI and cyber security	Identify a number of methods for analyzing big data besides describing the challenges facing the analysis process.

As has been shown above, big data has brought about a boom in the fields discussed above; the data has been growing at a faster rate which makes frameworks like Hadoop Spark and others useless. Therefore, it is necessary to make some adjustments or to create a newer, more modern framework whose capability is far superior to that of these frameworks in order for big data to be effectively analyzed, cleaned, processed, and used in the decision-making process. To be more precise, besides being efficient, a framework may be developed like Hadoop.

III. AN OVERVIEW OF BIG DATA

A. Big Data

Health care, financial services, and business recommendations are just a few of the many fields where Big Data is proving useful. According to The Economist, data is quickly becoming a crucial commodity for businesses. Input to the economy is comparable to both capital and labour. Today's data sets are not only massive in size and variety but also continuously changing. Data sources like WhatsApp, Twitter, Facebook, YouTube, and GPS signals from mobile devices are just a few examples. Consequently, Big Data is characterized by characteristics such as heterogeneity, lack of organization, semi-structure, incompleteness, and huge dimensions [15].

The word "big data" was first used in the 2000s through industrial data analyst Doug Laney, who outlined it in terms of the "three Vs" [16] [17]:

1) Volume: Companies get information from an extensive range of channels, like financial transactions, social media, besides sensor or device data.

2) Velocity: Unparalleled data streams are being received, and they need to be distributed properly. The development of the Internet of Things (IoT) sensors, RFID tags, with smart meters has raised the importance of managing data in real-time.

3) Variety: The term "data" refers to information in any form, from the numerical records found in traditional databases to the free-form text of emails, papers, videos, audio recordings, shares of stock, and monetary transactions.

All of the main IT businesses, such as Google, Amazon, Facebook, and so on, have begun big data programs. The availability of sufficient computing resources, analytical tools, and knowledge is essential for extracting information or data from large data. Making effective use of big data, therefore, requires producing value in relation to big data's scope, diversity, besides reliability [18].

B. Big Data Analytics Operations

It is possible to dissect the operation of knowledge discovery in databases (KDD) into its component pieces, as shown in Figure 1. In this context, "aspects" refers to the steps of choosing, preprocessing, transforming, data mining, and interpreting. If you follow the steps above, you'll have a fully functional data analytics system that can gather data, derive insights from the data, and show the results to the user [19].

In order to provide consumers with new insights, businesses engage in what is known as "data processing,"

which involves collecting, analyzing, and organizing data that already exists. Acquisition, Assembly, Analysis, and Action are the four divisions into which Karmasphere currently divides Big Data analysis. The 4 A's serve to highlight these measures [20].

1) **Acquisition:** The architecture of Big Data requires rapid data collection from numerous sources using a variety of controlled entry systems. It's where a filter might be located to retain just possibly useful data or at least partial data with reduced error probability. It may be helpful to collect such metadata and maintain it synchronized with the data itself for later analysis [21] since, in some applications, the context in which data was created is vital.
2) **Assembly:** The architecture must now be able to take in data in a wide variety of formats and derive meaning from it in the form of named entities, relationships, etc. Data must be made computable, structured or semi-structured, integrated, and stored effectively before they can be used in a computation. For this reason, we had to employ a version of Extract, Transform, and Load. There is no ironclad guarantee that cleaning will go easily within a Big Data architecture. In reality, we might not have enough time to completely cleanse all of Big Data because of its volume, velocity, diversity, in addition, variability [21].
3) **Analyze:** Here, we use querying databases, modelling data, and using computational methods to get a fresh understanding. Integrating, cleaning, and validating data are prerequisites for mining. In addition to gaining a deeper understanding of the data's semantics and enabling more insightful querying, data mining may be used to increase the data's quality and reliability [21].
4) **Action:** Accurately evaluating analytical results is crucial for making sound choices. As a result, it is crucial for the user to comprehend and validate results [21]. In addition, the source of the information should be disclosed so that the user may better understand the accuracy of the results.

C. Infrastructure for Big Data Analytics

Figure 1 below demonstrates the different layers that can be found in big data analytics.

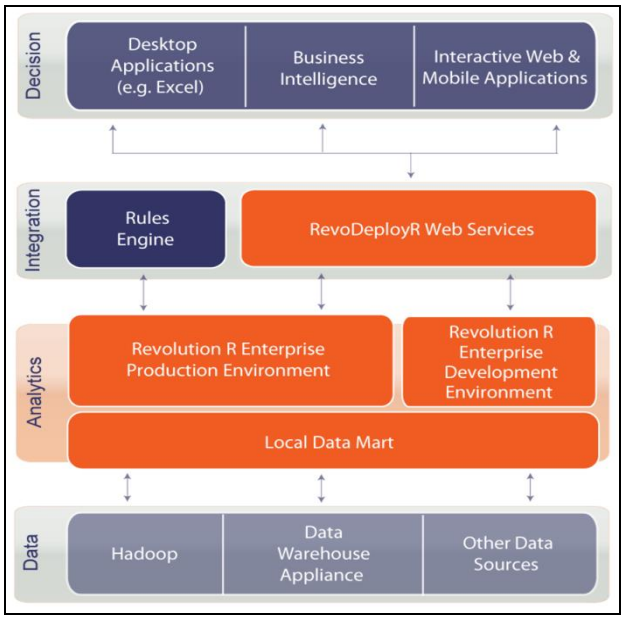

Fig.1 Architecture layers of big data analytics [20]

Here are the architecture layers [22]:

1) **Data Layer:** Here, we have data that is both structured and unstructured, including data stowed in relational database management systems. Unstructured data is best kept in NoSQL databases. NoSQL systems, as opposed to RDBMSs, offer alternatives like MongoDB and Cassandra. The term "unstructured" or "semi-structured" data describes a wide diversity of data formats, comprising but not limited to data streams from the Internet and social media, data from the Internet of Things sensors, and data from operational systems. At this stage, you'll also find software tools like HBase, Hive, Spark, and Storm. Even Hadoop and MapReduce depend on this layer.
2) **Analytics Layer:** lays the groundwork for analyzing data in real-time and making use of dynamic insights. It's set up with everything you need to create models, and it updates the local data on a regular basis. This boosts the analytical engine's efficiency as well.
3) **Integration Layer:** This layer connects the analytical engine to the apps used by end users. For real-time data processing, a rules engine and API are often required.
4) **Decision Layer:** In this stage, the finished product is released to consumers. Mobile apps, desktop programs, dynamic websites, and BI programs are all examples of end-user applications. Users will interact with the system at this tier.

Each of the aforementioned layers enables a vital stage in the deployment of real-time data analytics and is associated with a unique group of end users at the moment.

D. Comparison between traditional data analysis and big data analysis

There are two different ways to analyze data: traditional data analysis and big data analysis. Some of the main differences between the two are as follows:

1) Data size: Eternal data analysis works with extremely large datasets that may be too enormous for regular data analysis tools to handle, whereas traditional data analysis normally deals with small to medium-sized datasets.
2) Data structure: Structured data that is arranged in tables, spreadsheets, or databases is the focus of traditional data analysis. Big data analysis, in contrast, works with both structured and unstructured data, including text, photos, and social media data.
3) Processing speed: A single processor or a small cluster of processors can be used to do traditional data analysis. Contrarily, big data analysis requires distributed computing frameworks like Hadoop or Spark or parallel processing on numerous CPUs.
4) Analytical tools: Big data analysis requires specialist tools like Hadoop, Spark, or NoSQL databases, whereas traditional data analysis often employs statistical software like SPSS or SAS.

The entire comparison between traditional data analysis and big data analysis can be depicted in Figure 2.

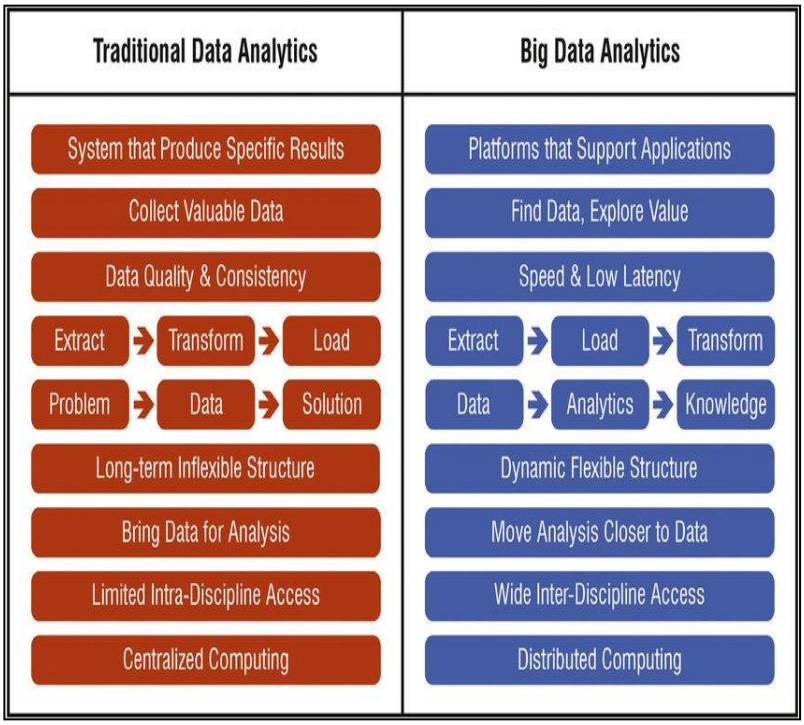

Fig.2 Comparison between traditional data analytics and big data analytics [18]

IV. TECHNOLOGIES FOR BIG DATA

Big data management is the operation of handling massive quantities of information, whether they are structured, semi-structured, or unstructured. When properly managed, big data can be used in a variety of business intelligence and big data analytics applications. To collect, store, and present massive amounts of data, there exists a wide variety of technologies for managing this data. This part talks about these tools and others like them. Here are a few examples of tools that are used for different things [23]:

A. Big data analysis frameworks and platforms

Systems and frameworks for big data analysis are readily available on the market. Below is a list of a few of the more typical ones:

1) Apache Hadoop: a platform that enables the distributed processing of huge data collections across computer clusters and is open-source.
2) Apache Spark: a big data processing engine that is open source and covers batch, stream, and machine learning workloads.
3) Apache Flink: a framework for stream processing that is free and supports both batch and real-time data streaming.
4) Apache Storm: a distributed real-time computing platform that allows for the quick processing of streaming data and is free.
5) Apache Cassandra: an open-source distributed NoSQL database management system designed to manage huge data volumes across many computers.

The popular big data technologies can be shown in Figure 3.

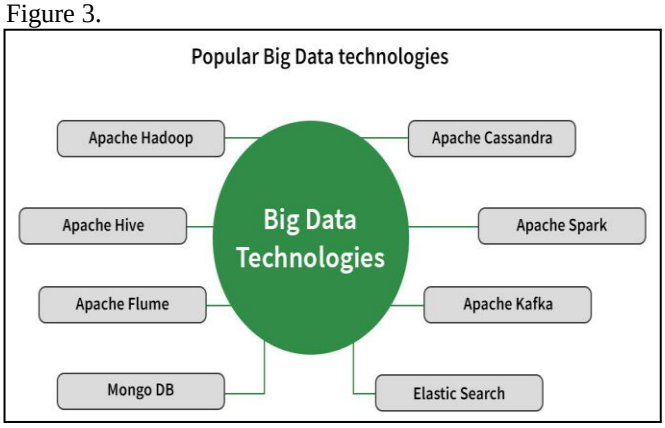

Fig.3 Big data technologies [17]

B. Storage Technologies

With the exponential increase in data sizes comes the requirement for reliable and powerful data storage methods. Technology related to data compression and storage virtualization has been the primary driver of progress in this area [29].

1) HBase

Deployable NoSQL database that sits atop HDFS. Apache HBase is a freely available open-source NoSQL database system that provides users with rapid access to very large datasets. In order to manage extremely big data sets, HBase can easily combine data sources that use different architectures and schemas, and it scales linearly to several billion rows and millions of columns. HBase is compatible with both Hadoop and the YARN access technique [30].

2) SkyTree

The primary application of this platform is in the field of big data analytics and management, where it serves as a high-performance data analytics and machine learning tool. Due to the massive amounts of data involved, manual exploration is impractical, making machine learning an essential component of big data. It is impractical to utilize automated methods of data exploration due to their high cost [31].

3) Non- Relational Databases

It is a way to manage data and build databases that work well for large amounts of data that are spread out over a large area. Apache Cassandra is used to build the most-used NoSQL database. Cassandra was Facebook's private database until 2008. In 2008, it was made available to everyone as open source. In the social media realm, NoSQL is used by Netflix, LinkedIn, and Twitter [32].

C. Equipment for Visualization

The market is flooded with open-source visualization tools. Only a small handful are included below [33].

1) R-Tool

To visualize data using statistical and graphical computing, many people turn to R, a language and software package widely used for this purpose. The R Project Statistical Computing backs this effort. Statistics professionals and data miners alike rely heavily on R to create and analyze statistical applications [34].

2) Tableau

Tableau is the software used to make visual representations of the data in the form of maps, charts, and graphs. Visual analytics can be performed with a desktop application [35].

3) Infographic

You can choose from a wide variety of visual templates, customize your presentation with additional visualizations like charts, maps, and movies, and then share your work with the world in just three simple steps. Accounts for classroom use and branding policies for enterprises are also included, as accounts for the publication of video and audio files and research scripts by journalists [36].

4) **ChartBlocks**

It's a free web application that makes creating visualizations from data sources like spreadsheets and databases simple and straightforward [37].

E. Comparison between the frameworks/platforms of big data

For large-scale data processing and analytics, there are several frameworks and platforms. Each has advantages and disadvantages of its own. Here is a brief comparison of a few well-known big data platforms and frameworks:

Apache Hadoop: Hadoop is an open-source system for the clustered processing and distributed storage of huge data collections. Hadoop Distributed File System (HDFS) for storage and MapReduce for processing make up its two primary parts.

Apache Spark: A cluster of affordable hardware is used in a distributed storage and processing system called Hadoop to handle massive data collections. HDFS (Hadoop Distributed File System), which handles storage, and MapReduce, which handles processing, make up its two primary parts.

Apache Flink: Flink is a framework for stream processing that allows for real-time processing of data streams.
Apache Cassandra: A distributed NoSQL database called Cassandra is made to manage enormous volumes of data across numerous commodity machines.

Apache Kafka: A publish-subscribe messaging system is offered by the distributed streaming platform Kafka. It has capabilities like fault tolerance, scalability, and high throughput and is built to handle real-time data inputs.
In conclusion, each of these frameworks and platforms offers distinct advantages of its own. In addition to the intended processing speed, fault tolerance, and scalability, the choice is based on the unique use case and requirements, such as the volume, velocity, and variety of data.

V. THE IMPOSSIBLE CHALLENGES OF BIG DATA

There are a lot of important things to keep in mind when working with Big Data and its analysis [38]. In their research, many people who study big data look at the following problems. Here's what they are:

- **Storage:** The size of hard drives in computers today is somewhere between terabytes (TB). Data production on the internet is measured in exabytes (EB). Though educational data creation isn't on par with that of the internet, it nonetheless generates a significant volume of

information. It will expand dramatically in the years to come. So, because this type of Big Data isn't structured, typical RDBMS systems like Oracle and MySQL can't store or process it. Cassandra and MongoDB are two examples of NoSQL databases that are used to solve this issue [39].

- **Data representation:** There are different levels of structure, meaning, type, organization, granularity, and access to many datasets. The value of information gathered through data analytics and user analysis can be improved with the aid of efficient data representation. If data are presented inaccurately, they may lose some of their intrinsic value and make in-depth analysis more challenging [40]. So, if the data is shown well, it will be easier to figure out what it means.
- **Management of data across its whole lifecycle:** In the analysis phase, the data life cycle management process chooses which data should be kept and which should be destroyed. There are problems, like the fact that the storage system we have now couldn't handle so much data. So, the life cycle management system needs a principle that makes it work well [41].
- **Analysis:** Big data is made by many kinds of online education websites, and each has its own structure and amount of data. Data analysis could be a costly and time-consuming process. To address this issue, "scaled-out" architectures are utilized to distribute data processing. Data are broken up into pieces and processed on different computers in a network. The processed data are then put back together [42].
- **Reporting:** A statistical report is a numerical representation of a statistical study. When dealing with massive amounts of data, conventional reporting approaches become complicated. In such situations, statistical reports call for a certain presentation format that is both comprehensible and concise [43].
- **Redundancy Reduction and Data Compression:** By eliminating unnecessary duplicates and compressing data, we can cut down on the system's overhead and save money. Data generated by sensor-based networks, for instance, is extremely redundant. It's possible to sort this information and cut down on duplicates [44].
- **Confidentiality:** Since service providers and data owners are unable to efficiently manage and analyze such massive datasets, data privacy is another major setback for big data. They rely on experts or external tools to examine this kind of data, which raises the stakes when it comes to possible danger. Therefore, protecting the privacy of the study data is a major concern [45].

VI. BIG DATA DIMENSIONAL REDUCTION

Dimensionality reduction is a technique that is essential for visualization, which is a vital part of microbiome data analysis. Several studies have recently focused on dimensionality reduction problems in large datasets. The following are descriptions of some of the work that has been done [46]:

The process of obtaining a collection of uncorrelated principal variables from a large number of independent variables is called dimensionality reduction. Features are either selected or obtained, relying on the objective. We could get more accurate plots and visualizations if we could reduce the data's dimensions to 2D or 3D. It's useful for compressing data and saving space. It reduces the amount of time needed to repeatedly execute the same computations. KNIME claims seven techniques can be used to minimize input dimensions. The following are the examples [47]:

1) Deleting columns from data sets that have no values

The data column will be unable to provide any information if there are any empty spaces where it should be. The purpose of this action is to get rid of the empty or missing data column. Find out how many data columns have missing values and then delete them [48].

2) Low variance Mesh

Data columns whose variance is below a specified threshold are filtered out by the Low Variance Filter node. Only columns containing numbers can be used in a variance calculation [48].

3) Reducing highly associated columns

Assuming that the values in one data column are substantially correlated with those in another, it may be

possible to eliminate the redundant columns without significantly reducing the amount of data accessible for future endeavours. Data columns that are too closely related to others can be ignored with the help of a linear Correlation node. One of two associated data columns can be discarded with the help of the Correlation Filter node [48].

4) Principal Component Analysis (PCA)

Using principal component analysis, we can reshape the x-coordinates of a dataset (PCA). In particular, PCA makes use of an orthogonal transformation [48].

5) The Forward-Facing Feature-Building Process

To create a collection of preselected classifiers, the forward feature creation technique uses an expanding set of input features. The method relies on a classification algorithm that gradually builds on a foundation of a single characteristic by adding others in successive iterations [48].

VII.ALGORITHMS FOR ANALYSIS OF BIG DATA

Data mining techniques and associated approaches to data analysis are crucial to big data analytics due to their importance in dimensionality reduction, memory demand and management, besides the correctness of the final findings. Here is a quick synopsis from the point of view of analysis and search algorithms that ought to give you some idea of its significance [49] [50].

A. Algorithm for Clustering

In cloud computing, CloudVista is one of the most often used clustering technologies for enacting the clustering procedure in parallel. CloudVista demonstrates its ability to manage massive datasets by employing BIRCH and other clustering techniques. One further technique for improving the efficiency and security of a clustering algorithm is the graphics processing unit (GPU) [49].

B. Algorithm for Classification

Classification algorithms [48] are similar to clustering algorithms in that they both take into account the input data acquired from the data sources and are controlled by a diverse set of learners.

C. K-Means

In order to cluster similar data or objects more, the k-means algorithm divides them into groups of varying sizes. It's a common method for organizing and analyzing large amounts of data [48].

D. PageRank

PageRank is another method of analysis that aims to standardize the relative value of a number of objects connected within a network of data objects. One goal of this method is to execute a certain kind of network analysis that seeks to discover connections between things and then rate them [49].

E. AdaBoost

A classifier is built by the Adaboost algorithm. A classifier is a tool that uses existing information to make educated guesses about which category a new piece of information should be filed under. In order to achieve its goal, this method combines several weak learners into a single robust one [49].

The algorithms of big data analysis can be described in Figure 4.

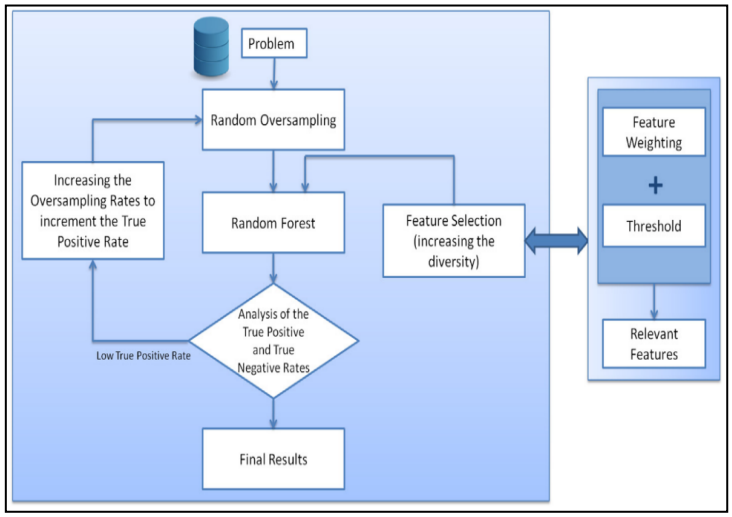

Fig.4 Big data analysis algorithms [48]

VIII. MODEL THE MINING PROBLEM TO FIND SOMETHING FROM BIG DATA

Data mining, which entails utilizing statistical and computational methods to examine massive datasets in order to find patterns, trends, and relationships, is the process of discovering something from big data.

You must first define the problem statement and the analysis's goals before you can model a mining problem. This can entail defining a specific business issue or goal, such as forecasting customer behaviour or spotting fraud.

Choosing and preprocessing the pertinent data comes after the problem has been identified. To ensure that the data is in an appropriate format for analysis, this may entail cleaning and changing it.

Choosing an appropriate algorithm or approach to analyze the data and produce insights comes after choosing the data. Classification, clustering, association rule mining, and regression analysis are just a few of the many distinct algorithms and methods that exist.

After the model has been created, it is crucial to assess its effectiveness and reliability. This may entail assessing the model's propensity to forecast outcomes or identify patterns in the data using metrics like accuracy, precision, recall, or F1 score.

Eventually, the analysis' findings must be accurately assessed and communicated to relevant parties, including company leaders and decision-makers.

IX. ANALYZING BIG DATA OF A SOCIAL NETWORK

There are many phases involved in analyzing social network data, including data collection, data preparation, data analysis, and data visualization. Here are some broad pointers for social network data analysis:

1. Define the research question: It is crucial to establish the study question before beginning to examine social network data. What are you hoping to accomplish with the analysis?
2. Collect data: There are several methods for gathering information from social networks, including web scraping, APIs, and data dumps. It is crucial to ensure that the data collection is moral and that the social network platform's terms and conditions are followed.
3. Preprocess data: Cleaning, filtering, and converting the raw data into an analyzer-friendly format are all parts of preprocessing data. In this stage, duplicates are eliminated, missing values are handled, and the data is transformed into a structured format.
4. Analyze data: Utilizing statistical and machine learning methods, social network data analysis aims to gain insights from the data. In this step, relevant traits are found, user behaviour is predicted, and similar people are grouped together.
5. Visualize data: Social network data can be shown to help convey the analysis's key findings. Creating graphs, charts, and other visualizations is part of this stage.

X. THE ROLE OF SECURITY AND PRIVACY ISSUES IN DATA ANALYSIS

Significant research difficulties in data analysis involve security and privacy concerns. It's critical to ensure that sensitive data is properly protected, given the growing volume of data being generated and gathered.

Large volumes of sensitive personal data, such as financial information, health information, and personal preferences, are typically collected, stored, and processed as part of data analysis processes. As a result, confidentiality and security issues may arise, especially if this data is compromised.

Data collection, storage, processing, and sharing are just a few of the many places where data analysis can lead to security and privacy issues. To prevent unauthorized access or data violations, it is crucial to verify that data is accurately safeguarded at every stage of the data analysis process.

XI. Conclusion

Several facets of big data, including big data analytics, big data analytics methods, data visualization, and the big data analysis algorithm, have been the subject of this literature review. A summary of the possible advantages of the big data research ecosystem is also provided by this survey. They are listed below:

• Scheduling algorithms are used to control the cloud-based platform's computer resources so that the data analysis task can be finished as rapidly as possible.

• Data privacy and data security, two connected but different issues, are inherited inquiry topics that offer instructions on how to keep and update data safely, ensure that data transmissions are encrypted, and prevent unauthorized parties from obtaining personal information. Despite the advent of the big data era, many problems regarding data security and confidentiality remain substantially intact from the days of conventional data analysis. Data security research will, therefore, unavoidably be involved in big data analytics research.

• A big data analyst's main duty is to shorten the time spent on input-related tasks, including comparing, sampling, and processing, in order to accelerate the analysis of massive datasets.

REFERENCES

[1] C. L. P. Chen and C. Zhang, “and technologies : A survey on Big Data,” vol. 275, no. 1, pp. 314–347, 2014.

[2] S. Jiang, X. Qian, T. Mei, and Y. Fu, “Personalized Travel Sequence Recommendation on Multi-Source Big Social Media,” IEEE Trans. Big Data, vol. 2, no. 1, pp. 43–56, 2016, doi: 10.1109/tbdata.2016.2541160.

[3] V. Dhoot, S. Gawande, and P. Kanawade, “Efficient Dimensionality Reduction for Big Data Using Clustering Technique,” vol. 1, no. 5, pp. 26–29, 2016.

[4] J. Archenaa and E. A. M. Anita, “A survey of big data analytics in healthcare and government,” Procedia Comput. Sci., vol. 50, pp. 408–413, 2015, doi: 10.1016/j.procs.2015.04.021.

[5] Yao Q, Tian Y, Li PF, Tian LL, Qian YM, Li JS. Design and development of a medical big data processing system based on Hadoop. J Med Syst 2015;39:23.10.1007/s10916-015-0220-8.

[6] Hui He, Zhonghui Du, Weizhe Zhang ,Allen Chen2, “Optimization strategy of Hadoop small file storage for big data in healthcare” Springer Science+Business Media New York 2017

[7] MM Rathore, A Paul, A Ahmad, M Anisetti, “Hadoop-based Intelligent Care System (HICS): Analytical Approach for Big Data in IoT ” ACM Trans. Int. Technol. 2017, 18, 8. [Google Scholar].

[8] S. Kumar and M. Singh, "Big data analytics for the healthcare industry: impact, applications, and tools," Big Data Mining and Analytics, vol. 2, no. 1, pp. 48–57, 2018

[9] D. Geng, C. Zhang, C. Xia, X. Xia, Q. Liu, and X. Fu, "Big data-based improved data acquisition and storage system for designing industrial data platform," IEEE Access, vol. 7, pp. 44574–44582, 2019.

[10] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, "Big data in healthcare: management, analysis and future prospects," Journal of Big Data, vol. 6, no. 1, pp. 1–25, 2019.

[11] A. Jaiswal, V. K. Dwivedi, O. P. Yadav. "Big Data and its Analyzing Tools: A Perspective," in Proc. IEEE 6th ICACCS'20, 2020, pp. 560-565.

[12] Y. Benlachmi and M. L. Hasnaoui, Big data and spark: Comparison with Hadoop, in Proc. 2020 Fourth World Conf. Smart Trends in Systems, Security and Sustainability(WorldS4), London, UK, 2020, pp. 811–817.

[13] Khan M, Malviya " Big data approach for sentiment analysis of Twitter data using Hadoop framework and deep learning". In: 2020 International conference on emerging trends in ınformation technology and engineering (ic-ETITE). IEEE, pp 1–5.

[14] L. Wang and C. A. Alexander, ''Big data analytics in medical engineering and healthcare : Methods, advances and challenges,'' J. Med. Eng. Technol., vol. 44, no. 6, pp. 267–283, Aug. 2020

[15] M. Gheisari, G. Wang, and M. Z. A. Bhuiyan, "A Survey on Deep Learning in Big Data," Proc. - 2017 IEEE Int. Conf. Comput. Sci. Eng. IEEE/IFIP Int. Conf. Embed. Ubiquitous Comput. CSE EUC 2017, vol. 2, no. July, pp. 173–180, 2017, doi: 10.1109/CSE-EUC.2017.215.

[16] R. B. B. Santos and J. C. Graves, "Extracting chaos control parameters from time series analysis," J. Phys. Conf. Ser., vol. 285, no. 1, 2011, doi: 10.1088/1742-6596/285/1/012002.

[17] C. Kacfah Emani, N. Cullot, and C. Nicolle, "Understandable Big Data: A survey," Comput. Sci. Rev., vol. 17, pp. 70–81, 2015, doi: 10.1016/j.cosrev.2015.05.002.

[18] A. I. Naimi and D. J. Westreich, "Big Data: A Revolution That Will Transform How We Live, Work, and Think," Am. J. Epidemiol., vol. 179, no. 9, pp. 1143–1144, 2014, doi: 10.1093/aje/kwu085.

[19] T. Hendrickx, B. Cule, P. Meysman, S. Naulaerts, K. Laukens, and B. Goethals, "Mining association rules in graphs based on frequent cohesive itemsets," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9078, no. 3, pp. 637–648, 2015, doi: 10.1007/978-3-319-18032-8_50.

[20] K. Sin and L. Muthu, ""Application of Big Data in Education Data Mining and Learning Analytics – a Literature Review "," ICTACT J. Soft Comput., vol. 05, no. 04, pp. 1035–1049, 2015, doi: 10.21917/ijsc.2015.0145.

[21] G. Manikandan and S. Abirami, "Big Data Layers and Analytics: A Survey," Lect. Notes Networks Syst., vol. 5, pp. 383–393, 2017, doi: 10.1007/978-981-10-3226-4_39.

[22] K. Davis and D. Patterson, Ethics of Big Data: Balancing Risk and Innovation. 2012.

[23] C. Ranger, R. Raghuraman, A. Penmetsa, G. Bradski, and C. Kozyrakis, "Evaluating MapReduce for multi-core and multiprocessor systems," Proc. - Int. Symp. High-Performance Comput. Archit., pp. 13–24, 2007, doi: 10.1109/HPCA.2007.346181.

[24] R. Mutharaju, F. Maier, and P. Hitzler, "A MapReduce algorithm for SC," in 23rd International Workshop on Description Logics DL2010, 2010, vol. 456.

[25] A. Thusoo et al., "Hive-a petabyte-scale data warehouse using Hadoop," in 2010 IEEE 26th international conference on data engineering (ICDE 2010), 2010, pp. 996–1005.

[26] A. Jain and V. Bhatnagar, "Crime data analysis using pig with Hadoop," Procedia Comput. Sci., vol. 78, pp. 571–578, 2016.

[27] Z. Prekopcsak, G. Makrai, T. Henk, and C. Gaspar-Papanek, "Radoop: Analyzing big data with rapidminer and Hadoop," in Proceedings of the 2nd RapidMiner community meeting and conference (RCOMM 2011), 2011, pp. 1–12.

[28] M. Strohbach, J. Daubert, H. Ravkin, and M. Lischka, "Big data storage," in New Horizons for a data-driven economy, Springer, Cham, 2016, pp. 119–141.

[29] D. Chrimes and H. Zamani, "Using distributed data over HBase in big data analytics platform for clinical services," Comput. Math. Methods Med., vol. 2017, 2017.

[30] R. Naqvi, T. R. Soomro, H. M. Alzoubi, T. M. Ghazal, and M. T. Alshurideh, "The nexus between big data and decision-making: A study of big data techniques and technologies," in The International Conference on Artificial Intelligence and Computer Vision, 2021, pp. 838–853.

[31] V. N. Gudivada, D. Rao, and V. V Raghavan, "NoSQL systems for big data management," in 2014 IEEE World Congress on services, 2014, pp. 190–197.

[32] N. Bikakis, "Big data visualization tools," arXiv Prepr. arXiv1801.08336, 2018.

[33] A. Malviya, A. Udhani, and S. Soni, "R-tool: Data analytic framework for big data," in 2016 Symposium on Colossal Data Analysis and Networking (CDAN), 2016, pp. 1–5.

[34] C. Rajeswari, D. Basu, and N. Maurya, "Comparative Study of Big Data Analytics Tools: R and Tableau," in IOP Conference Series:

Materials Science and Engineering, 2017, vol. 263, no. 4, p. 42052.

[35] K.-W. Su, C.-L. Liu, and Y.-W. Wang, "A principle of designing infographic for visualization representation of tourism social big data," J. Ambient Intell. Humans. Comput., pp. 1–21, 2018.

[36] S. K. A. Fahad and A. E. Yahya, "Big data visualization: allotting by r and python with GUI Tools," in 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE), 2018, pp. 1–8.

[37] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," Natl. Sci. Rev., vol. 1, no. 2, pp. 293–314, 2014.

[38] V. Marx, "The big challenges of big data," Nature, vol. 498, no. 7453, pp. 255–260, 2013.

[39] X. Jin, B. W. Wah, X. Cheng, and Y. Wang, "Significance and challenges of big data research," Big data Res., vol. 2, no. 2, pp. 59–64, 2015.

[40] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," J. big data, vol. 2, no. 1, pp. 1–21, 2015.

[41] D. Agrawal et al., "Challenges and opportunities with Big Data 2011-1," 2011.

[42] T. Huang, L. Lan, X. Fang, P. An, J. Min, and F. Wang, "Promises and challenges of big data computing in health sciences," Big Data Res., vol. 2, no. 1, pp. 2–11, 2015.

[43] F. A. La Sorte, C. A. Lepczyk, J. L. Burnett, A. H. Hurlbert, M. W. Tingley, and B. Zuckerberg, "Opportunities and challenges for big data ornithology," Condor Ornithol. Appl., vol. 120, no. 2, pp. 414–426, 2018.

[44] M. Naeem et al., "Trends and future perspective challenges in big data," in Advances in intelligent data analysis and applications, Springer, 2022, pp. 309–325.

[45] J. Wu, S. Guo, J. Li, and D. Zeng, "Big data meet green challenges: Greening big data," IEEE Syst. J., vol. 10, no. 3, pp. 873–887, 2016.

[46] M. Younas, "Research challenges of big data," Serv. Oriented Comput. Appl., vol. 13, no. 2, pp. 105–107, 2019.

[47] A. Fahad et al., "A survey of clustering algorithms for big data: Taxonomy and empirical analysis," IEEE Trans. Emerg. Top. Comput., vol. 2, no. 3, pp. 267–279, 2014.

[48] W. Xing and Y. Bei, "Medical health big data classification based on KNN classification algorithm," IEEE Access, vol. 8, pp. 28808–28819, 2019.

[49] A. Fernández, S. del Río, N. V Chawla, and F. Herrera, "An insight into imbalanced big data classification: outcomes and challenges," Complex Intell. Syst., vol. 3, no. 2, pp. 105–120, 2017.

[50] Pablo, R.-G. J., Roberto, D.-P., Victor, S.-U., Isabel, G.-R., Paul, C., and Elizabeth, O.-R. (2021). Big data in the healthcare system: a synergy with artificial intelligence and blockchain technology. Journal of Integrative Bioinformatics.