Journal of Computer Sciences

Original Research Paper

Transformative Effects of Big Data on Advanced Data Analytics: Open Issues and Critical Challenges

Shabnam Kumari and P. Muthulakshmi

Department of Computer Science, College of Science and Humanities,
SRM Institute of Science and Technology, Kattankulathur, Tamilnadu, India

Article history
Received: 26-01-2022
Revised: 06-05-2022
Accepted: 13-05-2022

Corresponding Author:
Shabnam Kumari
Department of Computer Science, College of Science and Humanities, SRM Institute of Science and Technology, Kattankulathur, Tamilnadu, India
Email: sk2581@srmist.edu.in

Abstract: We are living in the period of the information era (i.e., where data/information is the new oil for companies/industries). The term big data had been authored to depict this age. This study characterizes a lot on the idea of big data and its progress over the previous decade to till date. This study presents the meaning of this new idea and its qualities and additionally examines a few different things identified with few interesting issues. In big data, it is found that vulnerability can be taken care of physically or consequently (distinguishing proof or potentially separation of irregularities). Around 7 V's of big data, speed, gazetteers, and information bases should be persistently refreshed and information handled can be occasionally (at the standard time). It is found that in an event it is required to handle assortment, manage different information designs (tweets and regular language messages), and circulated information. To handle/tackle issues in sectors like medical science, agriculture, business, online trade, transportation, etc., big data role is essential. Smart things which are producing big data (through their communications) require analysis of this data for making effective solutions to such real-time applications. Hence, this study discusses work done towards big data concerning its current status, challenges to security and privacy, and future directions. In the last, this study may lead future researchers to a different potential of applications and research.

Keywords: Big Data Analytics, Privacy, Security, E-Healthcare, Web Mining, Machine Learning

Introduction

In actuality (physically), Data/big data doesn't exist. The interest in what is big data has been taking off in the previous few years. Here are some incredible realities given by Forbes, clients watch 4.15 million YouTube recordings, send 456,000 tweets on Twitter, post 46,740 photographs on Instagram and there are 510,000 remarks posted and 293,000 situations on Facebook. Simply, envision the colossal lump of information that is delivered with such exercises and this consistent production of information utilizing online media, business applications, telecom, and different spaces is prompting the arrangement of big data (Oussous *et al*., 2018; Yaqoob *et al*., 2016). In connection with this, the Internet of Things (IoT) addresses one of the principal markets of big data applications. Because of the high assortment of items, the uses of the Internet of Things (IoT) are ceaselessly advancing. In the present days, different big data applications uphold calculated endeavours, viz., vehicle tracking systems with sensors, wireless adapters and Global Positioning System (GPS), etc., and these applications are mostly driven by data enable organizations to monitor and manage their team (with optimized delivery routes) (Oussous *et al*., 2018). This is achieved through examining, exploiting, and combining various information gathered from the application including history. Nowadays, Smart City (particularly in India) is also a hyped research area based on the Internet of Things (IoTs) application.

It was (Cox and Ellsworth, 1997) used the term "big data" on massive data sets. Before finding "What is big data?" let us start with the knowledge of why the expression "big data" has acquired significance in the previous decade. Moving to this conversation, when was the last time, we utilized a floppy/a CD to store our information? If we are not off-base, do we utilize these gadgets in the mid-21st century? Presently, the utilization of manual paper records, documents, floppy disks, and so on, has gotten out of date. The purpose behind this is the outstanding development of information/improvement in innovation. Individuals started putting away their information in

Relational Database Systems. However, at this point, with the (hunger for) new creations, innovations, applications (required snappy reaction time with least/reasonable expenses), and the presentation of the web, it is lucid that even relational databases are deficient at this point. This age of nonstop and enormous information can be alluded to as "big data". While, data mining is the way towards applying explicit calculations for mining designs from this large information, i.e., the extra strides in the Knowledge Discovery Database (KDD) measure (Mishra and Tyagi, 2022; Mariscal *et al*., 2010), for example, information pre-handling, feature extraction, and determination joining or fitting earlier information and appropriate understanding of the after-effects of mining are fundamental to guarantee that helpful information has been gotten from the big data (see area 1.1.2 for additional subtleties). Besides this, in today's scenario getting quality information is a significant issue. The majority of the created/gathered data is unstructured, boisterous, ill-advised, and has no normalization. Along these, the associations/analysts need to apply extra exertion on this information to remove/change data into usable and important information for further use. In basic words, this error/inappropriate information should be cleaned for future exploration reasons.

Big Data: Definitions and Cycle

The most well-known fantasy related to big data is "It is just about the size or volume of information" Jagadish (2015). However, it was never just with "Big" measures of information (being gathered). In straightforward terms, big data alludes to the assortment of information from plenty of sources that can be dealt with and gotten to in a wide assortment of organizations (like content or media documents and so forth). In the previous decade (in the mid-21 century), there was gigantic information accessible (put away in a few data sets in a few areas), but since the fluctuated idea of this immense data, the customary social data set frameworks are unequipped for taking care of this colossal data. The three distinct types of information are, (i) Structured, which is organized with a fixed blueprint and addressed with lines and segments, for instance, RDBMS, (ii) Semi-Structured, which does not have a fixed arrangement, for instance, XML, JSON and (iii) Unstructured, which is an obscure pattern like audio, video, text records and so on. In the present-day situation, the storage and processing of big data prove to be a herculean task due to lacking advancements and the usage of primitive management devices and machines. Data foraging, proper data storage, easy and simplified data transactions along with scrutinized data analysis and capturing are a few obstacles faced in this field.

In the past decades, big data definitions have changed rapidly (and created confusion). Gartner Inc. (GIG) defines big data as, "big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making". Similarly, Tech America Foundation (TAFFBDC, 2012) defines big data as, "big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of the information". There are other several definitions of big data which had been proposed by several authors, but few can be found (Tyagi, 2019).

Types of Big Data

Big data can be categorized according to applications that are associated with e-commerce, agriculture, finance, e-healthcare, etc., (Gandomi and Haider, 2015; Emani *et al*., 2015). In healthcare, it can be classified into two categories:

i) Clinical Operations: Clinical operations refer to all the activities related to patients. It includes identification or diagnosing of disease, monitoring the progress of disease, trauma and observations, treatments, and reports about the patient treatments. The Electronic Healthcare Records (EHRs) (https://mohfw.gov.in/sites/default/files/7677310627EM.pdf) are useful for stratification and illuminating unknown correlations of diseases.
ii) Research and Development: The Research and Development (R and D) associated with health care are playing a vital role that dealing with big data. It is required to use statistical algorithms and tools to provide quality treatments and to increase the clinical trial design. Adaptive Clinical Trial (ACT) is to evaluate a medical treatment or device by the observation of patient outcomes and changing the parameters of the protocol trial in consensus with the previous observations and Predictive Modelling (PM) is to produce more effective drugs and devices. Apart from this, R and D related to big data can contribute to.
iii) Genomic Analytics: Execute gene analysis and gene sequencing as a part of the common decision process of medical care. It is found that the process is cost-effective and very efficient to identify changes in genes, associations with diseases and phenotypes, and identify potential drug targets.
iv) Evidence-Based Medicine: Analyse and combine a variety of unstructured and structured data like electronic medical care records, financial and operational data, genomic data, and clinical data to match treatments with better results, which in turn predict patients suffer from unsafe diseases and deliver them additional effective care.

For example, Patient Profile Analytics (Raghupathi and Raghupathi, 2014), Remote Monitoring and Accelerating Discovery. Hence in summary, today's big data can be depicted in pictorial shown in Fig. 1. The explanation of big data classification terms is discussed in Table 1 (refer to Appendix A).

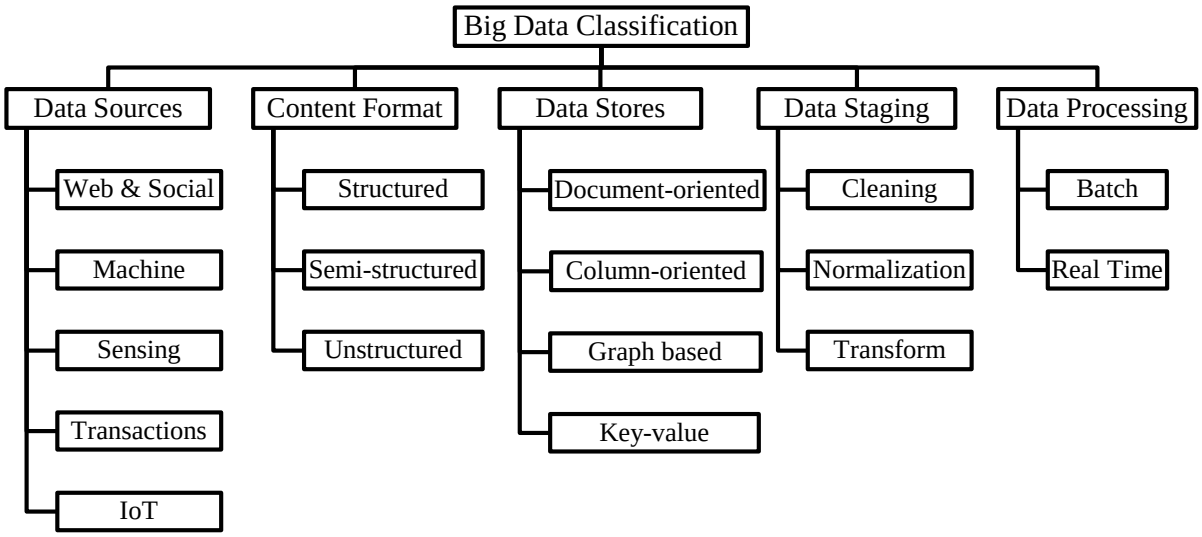

Fig. 1: Big data classification

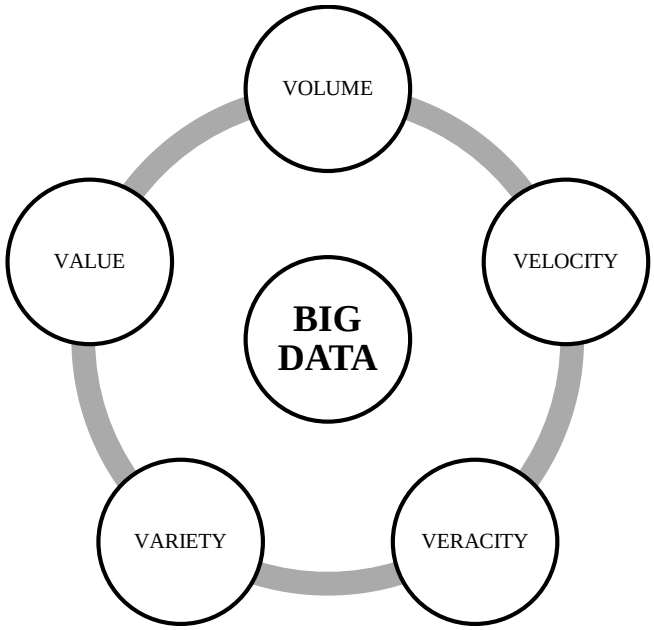

Fig. 2: Big data V's

Big Data Life Cycle

In general, the Big Data Life Cycle (BDLC) involves the states that find data from various sources through data delivery, the following are the process that represents the stages involved:

i) Collection of data: It involves the collection of data residing in different sources and storing them on to Hadoop Distributed File System (HDFS). The data collected can be of different formats, which may include medical image data, sensor data, social media data, etc.

ii) Data cleaning and pre-processing: It is an obvious fact that a majority of authentic data may not be consistent and hence cleansing of data helps in the eradication of missing data packs, repeated data values, noisy data layers, etc.

iii) Data classification: Segregating or sorting out data based on its structural outlook is involved in this stage

iv) Data modelling: The analytical survey done on the sorted and ordered data form comes under data modelling Considering the case where the government entails for a list of undernourished children in a given locality and this purpose, the data

has to be categorized based on the locality to which the children belong to after which the health report of the children are to be extracted which is followed by the identification of children belonging to families below the poverty line. In the end, all these details are to be processed, which involved data modelling data delivery: The most crucial step revolves around this stage, where triggering a report that underlies the data modelling stage which was carried out before this stage. As per the model examined, a report is delivered relating to the undernourished children of any area. This will assist the public authority to take important measures to evade further complexities. Big Data Life Cycle (BDLC) (Demchenko *et al*., 2014; Tsai *et al*., 2015) requires information stockpiling, data integrity, and control of information access at all stages.

In general, most data (big data) are of the unstructured form. To perform meaningful analysis, the data need to be classified into structured, semi-structured, and unstructured.

Background Work

The word ‘big data’ was coined numerous years back. Previously (long term prior) the principal endeavour was famously known as "data blast" to evaluate the development rate in the volume of information. There are a few stages (in the set of experiences) in regards to the size of information that lead to the advancement of the possibility of "big data". In 1944, F. Rider gauges that American University libraries were multiplying in size at regular intervals (remembered for (Lesk, 1997), and on this development rate, F. Rider determined that the Yale Library in 2040 will have "roughly 0.2 billion volumes (note that number of chomps for one book, at that point, it involves around 10 MB, at that point 1 G = 109 B = 100 books, i.e., 0.2 billion volumes = 0.2*109*10 MB = 2*1015 B = 2 PB)”. Cox and Ellsworth (1997) noticed that "informational indexes are for the most part that is very enormous and are burdening the limits of primary memory, neighbourhood circle, and even far-off plate". That could have been called as "issue of big data". At the point when informational indexes don't fit in fundamental memory (in center/on neighbourhood circle), the most well-known arrangement is to secure more assets. Lesk (1997) was found to utilize the expression called "big data" and the author raised a question that stated, "What amount of data is there on the planet?”. Also, the author answered and the statement was seen as the one followed here, "There might be a couple of thousand Peta Bytes (PB) of data". Raised questions on "The amount of Information?" and this was to be sure the absolute first endeavor to measure the amount of recently made and credible subtleties/information on the planet. Consistently the information is put away in four unique structures paper, film, CD's and DVD's. Attractive studies have brought into the spotlight the way in 1999 that 1.5 exabytes of particular information were delivered on the planet. Additionally, it can likewise be demonstrated that 5 exabytes of information were delivered in 2002. Further (Doug, 2001) distributes an examination note named "3D Data Management: Controlling Data Volume, Velocity, and Variety". 10 years after the said distribution acquired distinction, the called "3V's" are the for the most part acknowledged components of huge information which characterize it as all the three elements of life, and different analysts (Gantz *et al*., 2007) assessed the measure of advanced information made in 2006, i.e., the world made 161 exabytes of information and they came to the resolution that the data added yearly to the computerized universe is expanding more than six overlays to 988 exabytes, or multiplying like clockwork (somewhere in the range of 2006 and 2010). The volume of digital information created per annum rose dramatically to a stature of 1227 exabytes in 2010, which further rose to 2837 every year until 2012.

In 2008, Cisco anticipated that "Internet Protocol (IP) traffic will be almost twofold at regular intervals through 2012" and that it will arrive at a large portion of zettabyte in 2012. As an outcome, their expectation was correct, i.e., Cisco's report on May 30, 2012, gauged that IP traffic in 2012 at simply over a very large portion of a zettabyte. Bohn and Short (2009) stated that "The amount Information? 2009 Report on American Consumers", which uncovers that in 2008, Americans devoured data for about 1.3 trillion h, a normal of very nearly 12 h of the day. Of late, Bohn, Short (Bohn and Short, 2012) distributed an article in 2012 with the title "Estimating Shopper Information?”, which additionally uncover that in 2010, information utilization was expanding at a unique rate each day. Boyd and Crawford (2012) proposed "Basic Questions for big data". They also define Big Data as “a cultural, technological and scholarly phenomenon that rests on the interplay of, (i) Technology that maximizes the computation power and algorithmic accuracy to gather, analyse, link and compare large data sets, (ii) Analysis that draws on large data sets to identify patterns to make economic, social, technical and legal claims, (iii) Mythology that expands as a widespread belief of “large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible with the aura of truth, objectivity, and accuracy”.

It is found that several researchers discussed the magnitude of big data in terms of volume as PB, EB, and even ZB (Note that 1 PB = 210 TB, 1 EB = 210 PB, and 1 ZB = 210 EB) (Cai and Zhu, 2015). The particular highlights for big data showed up as '5V', where the initial three V's (Volume, Velocity, and Variety, refer Fig. 2) are those represented in 2001, and the fourth and fifth Vs stand for veracity and value. In the current scenario, it is found that

there are 5 V's in big data, but these V's are likely to be an increase shortly like Validity (accuracy of data), Variability (varying data behaviour), Volatility (dynamic changes with time), Vulnerability (can be easily breached or attacked) and Visualization (viewing a resourceful platform for data usage). This section summarizes the investigations on several facts related to the background of big data.

Big Data Analytics

To great extent, Big Data Analytics (BDA) is utilized by a few associations/organizations (like Amazon, Flipkart, Google, YouTube, Alibaba, and so on) to encourage their development and improvement. These organizations utilize/applies different information mining calculations on this information (caught/gathered information) and they use it to deliver better dynamic features. In light of the necessities and requirements of foundations and associations, there are enumerable devices for big data that are prepared like Hadoop, Pig, Hive, Cassandra, Apache Spark, Kafka, and so forth. Big data analytics can be organized into three kinds (i) "Descriptive analytics- the simplest class of analytics that allows condensing big data into smaller, more useful nuggets of information, (ii) Predictive analytics-a next step up in data reduction, with utilizing a variety of statistical, modelling, data mining and machine learning techniques to study recent and historical data, for making predictions, (iii) Prescriptive analytics-a type of predictive analytics to predict/ extract meaningful information. It is observed that big data are worthless if it is present in a vacuum and it would become valuable when it generates good or accurate decisions. To get great dynamic data, associations/organizations need effective cycles to turn high volumes of quick and assorted information into significant structure/data. The general interaction of removing data from big data is talked in five phases (Gandomi and Haider, 2015). These five phases structure the two principle sub-measures, Data Management and Analytics. This exclusively includes each one of those systems and helps procedures to save and store the information to keep it prepared for recovery with the goal that it tends to be dissected without any problem. Analytics focuses on techniques and methodologies for obtaining knowledge from big data.

Importance of Big Data Analytics

In today's era, the big data applications have evolved into fun-filled activities (Movies and video clippings are made by Netflix and Amazon for their users), automated cars, education, automobile, etc. Previously, this recommendation/prediction (or identifying customer choice) is done using Big Data Analytics (BDA) tools on this huge data (through data processing). With the assistance of big data Analytics and web-based media, both private and government offices help separate the concealed personal conduct standards of individuals. Among all applications, this data/information (gathered from/in clinical documented/gadgets) is a lot of valuable (for forecast) for medical care offices. Here, information analysis helps remove important data/information about medical care's market patterns, i.e., wellbeing area expects an improvement by finding concealed examples from this gathered big measure of (medical services) information. In the end, the investigation measure makes improvements in the conveyance of clinical benefits to the patients. For instance, a patient's reaction to a medication will help the drug organizations on medication advancement. Again, the investigation of big data encourages drug organizations to give customized prescriptions to every single patient to guarantee better and quicker recuperation.

Above examination measures can't be dealt with by customary frameworks, i.e., basic programming or equipment can't deal with or oversee numerous assignments on an enormous information. Subsequently, Big Data Management Systems (BDMS) is needed for taking care of/overseeing various assignments (productively). Big data, the executive's frameworks end up being very helpful as they recover the specific and exact subtleties needed by the clients for their examination. City traffic is another field where big data can be utilized to a Panglossian view. Big data can find the part in a wide assortment of boundaries like traffic streams and could assist organizers by reducing them from blockage and gridlocks, guaranteeing managed and smooth traffic streams over networks. Besides, the subtleties and data recovered from mobiles, personal computers, and so on can be used to upgrade the administrations that can provide services to users. When this gets into the hands of retailers, they can misuse this opportunity to improve client encounters by sending offers, advertorials, and so on. Besides, banks and monetary organizations can get benefits from overseeing viably dangerous liquidity.

Tools Available for Big Data Analytics

As discussed in the previous sections, big data is helpful to generate meaningful information for the future/to make new strategies for the organization. Several tools are used in different real-world applications to analyse big data.

Text Analytics

Text Analytics (text mining) alludes to methods that remove data from printed information. Text-based information models (gathered by association) include social organization channels, messages, sites, online discussions, overview reactions, corporate archives, news and call focus logs. Text examination includes factual investigation, computational etymology, and Artificial

Intelligence. Text examination empowers organizations to change over enormous volumes of human-produced text into significant outlines. For instance, text investigation can be utilized to anticipate securities exchange dependent on data extricated from monetary news.

Audio Analytics

This method examines and removes data from unstructured sound information. The dialects are used for examination (discourse investigation). Sound examination and discourse investigation are frequently traded in this field of work. At present, client call focuses and medical services are the essential application regions of sound investigation. Call focuses use sound examination for effective investigation of thousands or millions of long stretches of recorded calls. These strategies are used to improve client experience (with client conduct) in the future, likewise used to upgrade deals turnover, screen consistency with various arrangements (e.g., protection and security approaches), and recognize administration issues that are among numerous different assignments. In the future, audio investigation frameworks can be intended to break down a live call with or without Interactive Voice Response (IVR). Despite what might be expected, the sound examination ends up being incredibly useful in the clinical field as they back up analysis and therapy of intense and constant infections, which achieve sporadic correspondence results in patients' recovery procedure (e.g., despondency, schizophrenia, and malignancy and so forth).

Video Analytics

Video Content Analysis (VCA) advocates various techniques for managing, investigating, and recovering significant data from a grouping of videos. Nonetheless, video examination is viewed as a novel work in contrast with different applications presently being used. Before this, various methods were utilized for managing credible and pre-recorded recordings. The main consideration adding to the remarkable development of modernized video examination is the developing utilization of Closed-Circuit-Tele-Vision (CCTV) cameras. The significant test represented here is the humongous size of the information being used (i.e., 1 second of a superior quality video = 2000 pages of text). Also, consider the situation where 100 h of recordings are transferring at YouTube consistently. Here, big data advancements turn this test (of examining this enormous information) into a promising circumstance. The other key source added to the advancement of video examination is large information innovation.

Social Media Analytics

Social Media Analytics (SMA) is information that is organized and unstructured gathered from various diverse online stages. Web-based media envelops an excess of online sources allowing the clients to shape and exchange information. Web-based media stage is extensively seen as Social organizations (Facebook and LinkedIn), Blogs (Blogger and Word Press), Micro online journals (Twitter and Tumblr), Social news (Digg and Reddit), Social bookmarking (Delicious and Stumble Upon), Media sharing (Instagram and YouTube), Wikis (Wikipedia and Wikihow), Question-and-answer locales (Yahoo! Answers and Ask.com) and review sites (Yelp, TripAdvisor). Additionally, numerous versatile applications, (e.g., find my friend and so on) give a stage to social associations that fill in as web-based media channels with big data.

Predictive Analytics

The capacity to anticipate and foresee results sooner rather than later on the examination of provable information goes under prescient investigation. As a general rule, this is one such area which can be applied to all viewpoints of life, i.e., from anticipating a fly motor inability to speculating the following move of clients dependent on a point by point and customary report on their developments and decisions. Prescient investigation eliminates a few concealed examples and catches connections in this information. They are additionally characterized into two unique streams, (i) a few techniques that utilize information from the new and authentic past to foresee the result later, (ii) different utilizes recovering interdependencies between consistent result variable and illustrative factors. Relapse methods and arrangement, and grouping strategies are the two branches into which the scientific expectation procedure can be partitioned. Likewise, another group can be seen with strategies (in light of the kind of result factors), for instance, straight relapse tends to consistent result factors (e.g., to figure the deal cost of houses), while others, for example, Random Forests apply to discrete result factors (e.g., to decide credit status). In synopsis, classification is tied in with anticipating a name, and regression is tied in with foreseeing an amount.

Need for Big Data Analytics in Various Applications

Big Data is emerging as an opportunity for several organizations that are holding large data and these organizations use BDA for several benefits. Big data Analytics is to analyse large data sets to find all hidden patterns, unknown correlations (missing information), market trends, customer preferences, and other useful (business) information. Big data analytical tools are helping organizations improve their profits via effective marketing. Together with this, BDA is providing more (new) revenue opportunities to organizations, better services to customers, and improving operational efficiency and competitive advantages over rival organizations. Several industries use big data applications, which can be listed here as.

Manufacturing

Big Data proves to be highly advantageous in the manufacturing industries and similar fields concerning product quality, defects tracking, planning of supply distribution, differentiating between the defective and good products, increasing the efficiency of the devices and the machines, simulations, and check-tests of newly developed processes, etc.

Media and Entertainment

Furthermore, big data can be used to boost the status of entertainment industries as well, for example, by predicting the wants and expectations of the audience, optimizing the schedule, monetization of content, producing innovative projects and their implementation, etc.

Internet of Things

Extracted data from the Internet of Things (IoT) devices provides a mapping of device interconnectivity. This mapping is used by several various companies/organizations and governments to increase efficiency (or used to provide a happy and convenient experience/life to their users/ citizens). IoT is used to improve efficacy as the collected data from connected devices (using sensors) is used to analyse and take decisions at various levels in medical, agriculture and manufacturing, etc.

Government

The Government sector is using big data in several processes to increase the efficiencies in terms of functioning, cost, productivity, and innovation. The data sets (similar or different collected in government use cases) are used across multiple applications of government and it would increase the efficiency of government (when multiple departments collaborate and work together). The government can utilize big data to solve several problems in areas like Cyber Security and Intelligence, Crime Prediction and Prevention, Pharmaceutical Drug Evaluation, Scientific Research, Weather Forecasting, Tax Compliance (providing pensions to senior citizens without any delay), and Traffic Optimization (controlling traffic in peak times based on the live streaming data about vehicles).

Healthcare

The significant wellsprings of big data are web-based media locales, sensor organizations, advanced pictures/recordings, phones, buy exchange records, weblogs, clinical records, documents, military observation, internet business, complex logical exploration, etc., (as of now, the size of the gathered data is around some Quintillion bytes of information). This big data (gathered from a few sources/gadgets) should be examined for extricating meaningful data with innovation/apparatuses (with reasonable cost). This segment talks about the significance of dig data investigation in Healthcare. The administration of medical services applications can be increased by:

a) Providing patient-centric services*:* This refers to providing instant relief to the people suffering from various diseases and ailments by screening and detecting the presence of these harmful disease-causing microbes adjoined with providing the patients with a limited number of medicines at a lower dosage to tackle the advancing side-effects and rising prices of the medicines.
b) Detecting the spread of diseases at earlier stages: One of the major steps in preventing the occurrence of fatal diseases is to predict their presence at an earlier stage, this can be achieved with the help of the analytical data containing the social logs of the diseased in a particular location and it also gives the therapists and the doctors the necessary advice to cure the disease with suitable measures.
c) Monitoring the quality of the hospital: This step involves the consistent and strict monitoring of the hospitals to ensure that they are abiding by the necessary rules and norms set up by the Indian Medical Council. This also proves to be a useful step for the government to punish and disqualify the hospitals that fail to do so.
d) Improving the treatment strategies: Customized quiet treatment/observing the impact of the drug to be changed for quicker recuperation (in light of the investigation measurements of prescriptions). The patients can be monitored continuously and proactive care can be provided to respective patients. For example, analysis of data (produced by the patients who have already suffered from the same symptoms) helps doctors to provide effective medicines to new patients in future.

Note that some of the other benefits of BDA are, ensuring the reach of benefits provided by the government to all its citizens, mobile ambulance facilities, senior citizen care, orphan care, differently-abled care, women care, sanitary and cleanliness, waste management, e-waste management, etc.

Impact of Big Data Analytics on the Healthcare Applications

There are several benefits to the healthcare application of big data analytics, which can be summarized as:

i) Right Living: Patients can build value by taking an active role in their treatment, including disease prevention i.e., via having a proper diet and exercise

(i.e., when they get sick). The right-living pathway focuses on encouraging patients to make their lifestyle better and help them in being healthy.

ii) Right Care: This ensures that patients are getting appropriate treatment (i.e., the right care at the right time) available in a hospital. This can be ensured with the help of an interdisciplinary approach, i.e., all the staff and nurses should handle the same information working towards achieving a common goal.

iii) Right Provider: It provides well and affordable treatment by high-performing professionals to patients to achieve the best result. Here, 'Right provider' has two meanings, i.e., the right match of provider skill set to the complexity of the assignment (e.g., nurses or physicians' assistants performing tasks that do not require a doctor) and the cases where/we need a specific selection of professional doctors to provide the best result to patients.

iv) Right Value: Service providers and patients constantly upgrade medical services esteem through giving repayment to understanding results, or killing misrepresentation, waste, or maltreatment in the framework (while protecting or improving its quality).

v) Right Innovation: It includes recognizing and seeing novel treatments and thoughts to guarantee preeminent consideration and accommodation to the patients, i.e., boosting the R and D efficiency. To get the best outcomes, partners/organizations need to utilize earlier preliminary information, for example, by searching for high-likely targets and atoms in pharma (with discovering freedoms to improve clinical preliminaries and conventional treatment methods that utilize this information, particularly for births and inpatient medical procedures) (Peter Groves *et al*., 2013).

Impact of Big Data Analytics on the other Applications

This study can't be finished about information without discussing individuals who are getting profits from big data applications. Today, practically all the associations are utilizing big data applications in either way (i.e., in numerous territories). For instance, Telecom: Telecom areas gather data examinations it and give answers for various issues. By utilizing big data applications, telecom organizations have had the option to altogether decrease information bundle misfortune, which happens when organizations are over-burden and therefore the organizations will have consistent associations with their clients:

i) Retail: Retail has probably the most impenetrable edges and is perhaps the best recipient of enormous information. The magnificence of utilizing huge information in retail is to comprehend customer conduct, for instance, the profoundly productive proposal motor utilized by Amazon gives recommendations and assessments based on the perusing history of the buyers.

ii) Traffic Control: Traffic jams and congestions are a huge threat to several metropolitan cities all across the world (e.g., Delhi, Beijing, etc.) Technology-enabled sensors can be used in managing traffic better in big cities/highly dense cities (after analysing generated data).

iii) Manufacturing: Apprehension of big data in the industrial field seem to reduce a lot of defects in the number of components being made along with increased efficiency and therefore money and time are saved.

iv) Search Quality: Each time when a data is retrieved data from Google, a simultaneous data generation takes place based on which search quality is improved the next time the site is used.

This section discusses the need for big data analytics in various applications including the impact of using big data analytics in respective applications.

Machine Learning for Big Data Mining

And in the future, it is required to have more servers to store this large data, and for the organizations to run the business many more servers will be needed and the growth of server farming would be exponential. To consider the other side, analysis of the big data will lead to compressing the data without losing the meaning (after removing irrelevant/noisy data) and can be stored. The capability of Machine Learning (ML) for information examination can be effectively found in early writing (Elgendy and Elragal, 2014; Belle *et al*., 2015, Tyagi, 2019). In the previous years, information mining calculations (Mishra and Tyagi, 2022) is utilized/planned uniquely for explicit issues like classification, regression, and so forth, while the present AI calculations are utilized for mining diverse information (mathematical, ostensible, and so on). ML calculations can be utilized to locate an estimated answer for a streamlining issue. Consequently, ML calculations can be utilized for most information examination issues. AI is not just utilized for taking care of various mining issues in information examination but also in the administration of KDD (Knowledge Discovery Database) (Mishra and Tyagi, 2022; Mariscal *et al*., 2010). A new report (L'heureux *et al*., 2017) shows that AI calculations will be the fundamental piece of big data examination in the short future. It is noted that taking care of issues utilizing current AI calculations (for large information investigation) resembles conventional information-digging calculations. Subsequently, more consideration is required in

AI field to deliver productive strategies/calculations. Numerous issues can be tackled by means of "equal figuring" in AI, which is a form of Genetic Algorithm (GA) (Mansouri and Fox, 1991).

Data Mining

Data mining proves to be highly useful in hauling out the necessary raw data into productive information using ML and other statistical methods from the noisy data available.

The information mining procedures incorporate cluster examination, affiliation rule of learning, classification, and regression (Mishra and Tyagi, 2022), and the techniques are used for information recovery. The calculations in (Oyelade *et al*., 2016) talked about hierarchical clustering, k-means, and Fuzzy C-Means (FCM) used in grouping enormous applications. CLARANS (Clustering Large Applications dependent on Randomized Search) and adjusted iterative lessening and clustering utilizing chains of importance must be spread out for the prescient utilization of huge information clustering else, the current calculations would end up being a squandered dump. Current information mining calculations are not adequate to deal with the big data (due to the high rate of data/information creation) and it is very much required to have new and refreshed big data mining methods to take care of the huge heaps of information. Stream mining field is an illustration of continuous information mining that uses KD Nuggets to do the process and the following content could explain the potential areas that use continuous information mining.

Big Data Mining

Big Data is called for the need for real-time processing, data retrieval, and advancements in the field of decision support. Utilizing data mining techniques through the aid of cloud computing will facilitate the end-users to retrieve meaningful information from the integrated data store. Cloud computing technology reduces the cost of setting up the infrastructure for computing and storage and here are a few situations where cloud technology, big data, and data mining go hand in hand, Identifying and pulling out exceptions and encoded subtleties from big data of high volumes:

i) Mining geospatial and topological organizations and connections (utilizing AI and insight) from the information of IoT.
ii) Developing an all-focused methodology situated towards the appropriation of information mining calculations and rationales to the distributed computing hubs for information mining.
iii) Developing an elective sub-division of effective mining techniques that improve the capacity and preparing limits of cloud.
iv) Addressing the assortment of information mining difficulties by administering and surveying the presently utilized spatial mining techniques for its prosperity or disappointment.
v) Providing new mining calculations, apparatuses, and programming in the hybrid cloud administration frameworks.

Quantum Computing Based Mining

Shortly quantum computers will be used to mining big data efficiently. Through this process, energy, and time could be saved and predictions related to any problem like the stock market, weather prediction, cause of any disease or solution for a disease, etc., can be found easily. But as the biggest challenge, we require a sufficient number of skilled workforces to handle or perform such complex computing mechanisms. Note that Artificial Intelligence-based Data mining is a part of quantum computing-based data mining.

Web Mining

Web mining is a method used to identify a particular pattern from a reservoir of webs. It exposes the hidden knowledge of a given site and its users to carry out analysis and eases the analysis of the effectiveness of a particular website. Web mining is classified into two different types as follows:

i) Web content mining/Web mining: This is exceptionally helpful in recovering the needed and important information from a pool of web content which would comprise sound, video, text, and pictures. The different nature and disorderly structures require an exceptionally extended data source on the World Wide Web to give solace to the clients. However, these devices are not sufficiently grown to allow coordinated information designs and records ending up being a profoundly compelling variable for the exploration engineers to advance more shrewd devices for information extraction and to spread its data set to give an undeniable degree of authoritative incentive on the web (Khan *et al*., 2009). This would require the creation of an intricate and complex AI framework equipped for self-ruling action to find and structure web information (Xu *et al*., 2011).
ii) Web structure mining: This is implemented to capture the hub and associated structure of a site by realistic perception. Additionally, web structure mining is partitioned into two fundamental fields, (i) design extraction from hyperlinks inside a site and (ii) investigation of a tree-like construction to portray HTML (Hypertext Mark-up Language) or XML (extensible Mark-up Language) labels.

On another side of web mining, visualization is a process of analysing the data by seeing them through graphs, tables, and diagrams. It's a known and experienced fact that big data visualization is a herculean task when compared to small data visualization due to its intricacy of the four Vs (Geng *et al*., 2011).

Machine Learning

Computer behavior can be easily evolved using Machine Learning (ML). However, the existing ML techniques that are supervised and unsupervised have to be enhanced to cope with the bid of data processing and skeletal structures like Map/Reduce and DrvadLINO can be used for the same. The algorithms used for this approach are still similar to those of off-springs and possess scalability issues. Besides, Artificial Neural Network (ANN) is utilized in deep learning, pattern distinguishing proofs, versatile investigation, and others. Note that ANN is regularly used to address/investigate the huge scope of datasets. In basic words, the intricate learning interaction of ANN throughout big data is tedious. This is a serious concern/issue to consider in future work concerning ANN. It is recommended to refer work of (Pramod *et al*., 2021; Tyagi, 2019; Tyagi and Chahal, 2020) for knowing more (depth-knowledge) about machine learning and its subset learning techniques (deep learning).

Optimization Methods

These techniques are misused to conquer process-able issues and are utilized in an interdisciplinary field, for instance, they include an assortment of procedures to handle worldwide enhancement issues which incorporate mimicked strengthening, quantum tempering, swarm optimization, and genetic calculations (Li and Yao, 2011). Though they may kill up a lot of time and may be highly complex, they provide optimization to a high extent. By scaling up and enhancing these processes, they can be easily applied to big data and will be a big booster for optimization in this field too.

Social Network Analysis

Viewing social dealings and affairs in the network theory is eased with the help of Social Network Analysis which has gained a great deal of significance in cloud computation. Yet, Social Network Analysis (SNA) showcases decreasing performance when there is a huge amount of data and dimensions are in use and it is quite difficult to access this in the current-research world (Bingham and Mannila, 2001).

In summary, we can say that there is an exponential rise in the big data/information generated with the advancement in science and technology. As a result, there are numerous challenges related to the rapid growth of data. In this specific circumstance, this study is a thorough examination of cutting-edge (allude Table 1 and 2 in Appendix A) information investigation strategies that include information mining, web mining, AI, SNA, representation and improvement techniques, and so on.

Critical Challenges Towards Big Data and its Implementation

Truth be told, while the size of big data continues to increase dramatically (ordinary), the current mechanical perspectives/techniques neglect to deal with and investigate these enormous big Data sets (i.e., petabytes, exabytes, and zettabytes of information). And therefore, it is essential to have new machine learning methods/algorithms to solve the problems discussed so far. Apart from security and privacy issue (which need to be implemented in big data) (Nambiar *et al*., 2013), several other challenges are also notified in the area of big data and are listed:

i) How to team up, gather and store alongside the restricted equipment and programming gear, humongous information parcels from various sources.
ii) How to achieve the multifaceted nature of big data and cross it in an alternate climate with a blend of application utilization.
iii) How to collect enormous information that is faultless and how to distinguish the dependable one, which ends up being profitable.
iv) How do blend unfamiliar information sources and disseminated big data stages with the inward foundations of an association? However, it is not sufficient to break down the information (created inside associations). To expand the estimation of important information, it is fundamental the blended total inside information with outside information sources. Here, amassing inside information with outside information sources is a significant issue.
v) Note that the computer design and its ability due to the Central Processing Unit (CPU) execution, CPU/Processor size is multiplying its ability every year and a half (due to producing enormous information by gadgets), i.e., the exhibition of plate drives is additionally multiplying at a similar rate (according to Moore's law). Nonetheless, the I/O tasks do not follow a similar exhibition design (i.e., irregular I/O speeds have improved tolerably while consecutive I/O speeds increment with thickness gradually). This imbalanced framework limit may moderate getting to information and influence the presentation and the adaptability of big data applications. Again, it can be seen that different gadgets limits an organization (i.e., sensors, circles, recollections), i.e., this may hinder the execution of the framework.

i) Big data, the executives: It is exceptionally hard to deal with big data proficiently, i.e., in the extraction of solid and significant data (with enhanced costs). Coordinated and effective information is a firm help for BDA. Big data implies cleaning information for dependability, gathering and isolating information in the wake of amassing from a plenty of sources, and encoding the equivalent for protection and security issues.

ii) Distributed mining: As examined in segment 4, the majority of the investigation strategies don't work in an equal manner, i.e., these methods (customary information mining procedures) are incapable to dissect (mine) this enormous information. Consequently, it is required to create circulated variants of existing examination strategies (it will require ability and a ton of exploration) to dissect a lot of appropriated information in a proficient manner.

iii) Scalable Artificial Intelligence: To extricate important data from a huge information (i.e., uproarious information), it is required to have adaptable (productive and precise) AI calculations. The current ML codes and rationales are not executable for immense measures of information (Yaqoob *et al*., 2016). In the future, adaptable AI calculations are should have been utilized to tackle information adaptability issues.

iv) Communication cost: Communication cost happens between frameworks (of information examination) when correspondence is held. How do diminish the correspondence cost among frameworks? Additionally, another inquiry is "How the big data investigation speaks with different frameworks"? Consequently, significant inquiries are here, "How to diminish the expense of correspondence and how to make the correspondence solid between these frameworks"? These are the two significant open issues for big data examination.

v) Time Variable information: Data is changing so quick (in size and configuration) these days, so it is significant that large information investigation procedures (information mining/AI) should have the option to embrace these changes and distinguish these progressions.

vi) Mining from sparse data: This is a one-of-a-kind characteristic of big data applications (Yaqoob *et al*., 2016), however recovery of the solid end might be intense from inadequate information. The inquiry on the information measurement issues in big dimensional space does not exhibit information appropriation (making it hard to apply any mining procedures).

vii) Big data classification plans: When information (bigger information) is put away in a circulated way, at that point it is too hard to even think about recovering the required data (on/as expected). Subsequently, extraction of significant and vital information requires the presence of various classification calculations as the current ones were given remembering the extraction of information from limited quantities of putting away information. Even though various individuals have been abiding in this field of exploration, they are yet to bring out prosperous changes.

viii) Analytics/Data Analytics: Sheathing out helpful information like examples and rationales can permit associations in getting better and more astute. However, the extraction of required information stances to be a danger. In the 21st century, the current instruments can unsheathe the examples with a disadvantage of diminished exactness. Subsequently, future improvements should revolve around the examination of enormous and complex information.

ix) Data Quality: The primary utilization of information investigation is to gather and isolate information in an organized way for shopper prerequisites even though the information esteem for unequivocal control influences it when the nature of information is not the first-rate or if it contains commotion. Henceforth, we need to explore information quality administration issues to clean information with effective procedures (Mishra and Tyagi, 2022).

x) Visualization: Visualization alludes to addressing information (by utilizing charts). Yet, existing techniques/apparatuses do not give ideal execution time (concerning usefulness, versatility, and fast reaction time) for huge information representation. Rather than utilizing these (old) representation apparatuses, we need to reconsider "How to picture big data alternately is found essential "? and others like clamor, anomalies, deficient and conflicting information: As examined so far, big data examination is another age for information investigation or to remove significant or commotion free information. It is observed that the conventional information mining calculations sometimes fall short for current big data/new data set frameworks (created information from various sensors and frameworks). Here, open issues of clamor, anomalies, and deficient and conflicting information will be consistently present in big data mining calculations (like in conventional information mining calculations), but with high precision than the present-day strategies. Note that the impact of commotion and wrong and wasteful information will be upgraded for big data examination and therefore the question that states "How to relieve the effect?" will be an open issue for big data examination/future analysts.

xi) Security: The information (produced from a few gadgets/sensors) is gigantic in size and getting this is a tremendous undertaking as it incorporates client validation, restricted information access, recording

information access chronicles, legitimate information encoding, and so on.

xii) Classifying imbalanced dataset: This has been a hotly debated issue for various years. Talking, true applications are probably going to create various fields with dissipated sources, i.e., fields that are not first-rate and those which have various occasions. Along these lines, information researchers/specialists are welcome to put forth a strong effort (to tackle this issue) in this research field/area.

xiii) Others: Although Structured Query Language (SQL, a non-procedural language) information bases have a few points of interest, for example regarding adaptability, open-source, financially savvy and versatility, these data sets are experiencing a few issues (because of a lot of information) like the absence of development and consistency (identified with execution). Likewise, NoSQL (not just SQL) data sets do not manage the investigation. In this way, to exhibit constant information taking care of undeniable level figuring skeletal design and the undeniable frameworks to address and register logically.

xiv) Lack of Talent: Though there might be various venture advancements in the setup establishments, getting hold of scientists and staff who are gifted and solid in this field are elusive.

Here, the consistency of information between various frameworks, modules, and administrators is likewise a significant open issue (on the correspondence between frameworks). We need extra exploration to plan effective information recovery calculations (to remove important information) from a lot of information. Thus, from all difficulties/issues, we sum up that it has gotten extremely testing because of the intricacy and continuous handling requests of streaming information (created at colossal speed and evolving powerfully) to plan and execute new security systems (to ensure information and ID. of the malevolent information) that can secure the information immediately in the preparing.

Challenges in Big Data Management

In this segment, we examine the flow of research that highly focuses on the issue of Big Data Management (BDM) for investigation reasons. Nonetheless, there are many open provokes accessible concerning big data Analytics. The rundown talked about so far (counting underneath) is not conclusive and we require a ton of exploration work to do in the not-so-distant future.

a) Data Variety: How to deal with a continually expanding volume of information? Particularly when the information is unstructured or chaotic in structure and similarly how to separate significant data/content from it (in the least time and low cost)? How to total and associate streaming information from numerous sources.

b) Data Storage: How to distinguish and save intense subtleties recovered from chaotic information? How to capacitate voluminous data which can be extricated as and when required? Are the current frameworks streamlined according to the requirements? If not, what new augmentations are required? How could the put-away information be handily ported and moved without weakening.

c) Data Integration: New standards and guidelines for information mix that can deal with various kinds of information and sources.

d) Data Processing and Resource Management: Latest models which are enhanced for multi-dimensional information; new web indexes that can productively deal with improved document frameworks; motors fit for partner applications from various models (e.g., Map Reduce, work processes, and pack of-undertakings) on a solitary extraction.

e) The size of the information is continually increasing and is persistent. Soon, there is a likelihood that more V's will be presented in the not-so-distant future. Coincidentally, we can find in future 10 V's, i.e., volume, velocity, variety, variability, veracity, validity, vulnerability, volatility, visualization, and value. Consequently, with expanding of V's in big data, additional difficult issues will emerge in future.

Challenge in Big Data Based Cloud Environment

It is obvious that several smart devices (concerning real-time applications) are associated with cloud systems and huge data are generated every day. Now questions like "management of this data and control on this data and security and privacy, etc." are remaining unanswered.

Shockingly, very little work has been done/to make the information mining and delicate figuring calculations work on Hadoop. Individuals from a few unique foundations/future analysts need the plan and improvement of such undeniable level coherent calculations. Another common challenge that is faced is with the centralized system approaches and found that the majority of the developed algorithms have a centralized computing approach i.e., they are permitted to work on all the data simultaneously and because of this, designing a logical system that provides a channelized approach for allowing parallel computation is also beyond the scope. We need to think about the above-discussed issues/challenges when we work in the area of big data. Further, challenges in AI, ML, deep learning, etc., have been discussed by (Pramod *et al*., 2021; Tyagi, 2019; Mishra and Tyagi, 2022). Hence, this section investigates several open issues/challenges that need to consider in future work. The next section will provide a shadow on some critical issues like privacy, security, etc., and which are highly critical to solving.

Security and Privacy Issues Towards Big Data

In general, today's security and privacy are essential parts/components of any technology. When there are lots of devices (Internet of Things/Smart Devices) integrated into the cloud then it generates huge data every day. Hence, communication process, application development, security, and privacy issues are in high rise condition and serious attention towards these is very much appreciated.

Security Issues

It is seen everywhere that data (computing environment) processing and human are working together. Now, the question is "how to protect the data from humans" and it is an open issue to solve (gather more data from everywhere and keep it securely and then make analyses on it). Note that with a safe cycle (for gathered information), the big data investigation can be a dependable framework. From our new investigations, records, and examination, wellbeing issues of big data examination are four overlays: Input, information examination, yield, and correspondence with different frameworks. Concerning input, it is viewed as pertinent to information assortment and collection through the sensors, handheld gadgets, and even the frameworks and gadgets of IoT (also called as Web of Things). In general, the pursued point in security issues is identified with the versatility offered by the sensors to attacks. For the examination and information, a contributor to the issue, the key component is the security of the framework all in all. Also, last but not the least, the correspondence with different frameworks has a security issue raised between big data investigation and other outer frameworks. Because of the previously mentioned reasons, security has vanquished the other existing issue identified with BDA.

Privacy Issues

Privacy and Trust, have always been an area where a maximum number of attacks are mitigated in the current scenarios, especially when their details and information are at stake (or will not be used for another purpose but rather for it was stored/collected). The main question that arises in the minds of the people is highly genuine-on the off chance that the facts confirm that the frameworks can re-establish the individual data from enormous information investigation, however, the information might be mysterious. With protection issues being raised all over the globe in an assortment of fields, it has additionally grabbed an immense eye in information mining as the extricated private data and subtleties might be spilled to or vested in unlawful hands, for instance, albeit all the assembled information for shop conduct are unknown (e.g., purchasing a gun), because the information can be handily gathered by various gadgets and frameworks (e.g., area of the shop and age of the buyer), an information mining calculation can undoubtedly deduce the person who purchased this gun. Honestly talking, information investigation diminished the extent of the data set as the area of the shop and the age of the buyer gives the fundamental subtleties to help the gadget sort out potential individuals. For the same explanation, all unobtrusive data requires to be protected cautiously. The vague, transitory ID and encoding are the special and bleeding edge innovations for information security, yet the basic factor is the way, what, and for what reason to utilize the held information on big data investigation. Hence, the privacy issue is a hot issue and a much-needed issue (concern) to solve in the computing/digital world by future researchers.

Ethics and Privacy

Morals and protection consistently have been a significant worry in the information board. Presently, it has become an enormous interest in big data. There are several issues due to the multi-dimensional representation of data and a few are listed here:

- The presence of voluminous data, precious and intricate details and information can be recognized more than it was possible before.
- The valuable data velocity permits sufficient analysis and allows consistent refining of user's profiles.
- The huge collection of data from a spread of sources makes it easier to trace and track people along with an additional factor of providing data owners to develop sophisticated and enriched profiles of consumers.
- Security is not the authorization to hold privileged information yet ought to be viewed as an assortment of rules which licenses the stream of data in moral manners without suppressing one's opportunity. Nonetheless, shared data might be classified too.

Hence, the source of information from associations (clinics, drug stores, organizations, clinical focuses, and so on) is in various configurations. These associations have information in various frameworks with various settings (at a few spots). In the present situation, big data platforms are strengthened by various preparing and examining instruments along with the perception and creative mind that can recover subtleties and qualities from the intricate powerful environment. They additionally end up being tremendous assistance in dynamic by the suggestions and self-sufficient location of ambiguities and surprising conduct of latest things. This study attempts to consolidate the clarifications about the flow of research on big data. This study talks about big data attributes, the effect of large information investigation in different applications; also, it examines the difficulties raised by big data processing frameworks. This study broke down from the correlations made in (Yaqoob *et al*., 2016; Mishra and Tyagi, 2022; Gandomi and Haider, 2015, Raghupathi and Raghupathi 2014) that the

greater part of the current investigation strategies can function admirably for organized information as it were. Be that as it may, the greater part of the information is in unstructured and semi-organized configurations, which make a few difficulties for existing techniques/apparatuses. At that point, the open issues on calculation, nature of the final product, security, and protection are examined in detail in this study. Henceforth, this section talks about a few essential issues raised regarding the implementation of big data in many applications/ sectors. Though, we locate that big data mining requires straightforwardness because huge information can compromise protection. The next section will conclude this study in brief.

Conclusion

This study presents a top-to-the-bottom point of view into the design, strategies, and activities which are continued in big data computing (including issues and challenges). For thinking of imaginative forward leaps in the big data field, it is understood that (from our examination made of different advances in Table 1 and 2, in Appendix A) there are various obstacles and errands to conquer the vast majority of may be connected to form structures and systems. Consequently, in the coming future, territories like information association, space explicit instruments, and stage apparatuses should be centred around creating cutting-edge big data frameworks. Then again, from the viewpoint of the information mining issue, this study gives a concise prologue to the information and large information mining investigation apparatuses.

To more readily comprehend the progressions achieved by the enormous information, this study zeroed in on the information examination of KDD from the stage/structure to information mining. In outline, this research work audits a few major information logical methods for organized and unstructured information. Besides, processing big data/enormous information requires superior and versatile mining calculations/frameworks/instruments (which can work in equal and in disseminated design) to perform an investigation in an ongoing environment. With that, the primary advantages can be recognizing illnesses at prior stages, identifying medical care misuse and extortion quicker, and lessening costs. It can profit patients, researchers and developers, and healthcare providers. Consequently, innovative issues in numerous big data areas can be additionally considered and make it a significant examination theme to get focussed.

Acknowledgment

We, the authors want to thank all the Anonymous Reviewers, Publishers for their speedy response. We also thank to SRM Institute of Science and Technology, Kattankulathur, Chennai for their valuable support.

Authors' contributions

Shabnam Kumari: Written/drafted this study.

P. Muthulakshmi: Supervising, approved for final publication.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Lakshman, A., & Malik, P. (2011). The Apache cassandra project.

Abadi, D. J., Boncz, P. A., & Harizopoulos, S. (2009). Column-oriented database systems. Proceedings of the VLDB Endowment, 2(2), 1664-1665. doi.org/10.14778/1687553.1687625

Belle, A., Thiagarajan, R., Soroushmehr, S. M., Navidi, F., Beard, D. A., & Najarian, K. (2015). Big data analytics in healthcare. BioMed research international, 2015. doi.org/10.1155/2015/370194

Bingham, E., & Mannila, H. (2001, August). Random projection in dimensionality reduction: Applications to image and text data. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 245-250). doi.org/10.1145/502512.502546

Bohn, R., & Short, J. E. (2012). Info capacity| measuring consumer information. International Journal of Communication, 6, 21. https://ijoc.org/index.php/ijoc/article/view/1566

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological and scholarly phenomenon. Information, communication and society, 15(5), 662-679. doi.org/10.1080/1369118X.2012.678878

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. Data science journal, 14. doi.org/10.5334/dsj-2015-002

Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., ... & Gruber, R. E. (2008). Bigtable: A distributed storage system for structured data. ACM Transactions on Computer Systems (TOCS), 26(2), 1-26. doi.org/10.1145/1365815.1365816

Chen, Y., Alspaugh, S., & Katz, R. (2012). Interactive analytical processing in big data systems: A cross-industry study of mapreduce workloads. arXiv preprint arXiv:1208.4174. doi.org/10.48550/arXiv.1208.4174

Cox, M., & Ellsworth, D. (1997, October). Application-controlled demand paging for out-of-core visualization. In Proceedings. Visualization'97 (Cat. No. 97CB36155) (pp. 235-244). IEEE. doi.org/10.1109/VISUAL.1997.663888

Das, S., Agrawal, D., & El Abbadi, A. (2010, June). G-store: A scalable data store for transactional multi key access in the cloud. In Proceedings of the 1st ACM symposium on Cloud computing (pp. 163-174). doi.org/10.1145/1807128.1807157

DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., ... & Vogels, W. (2007). Dynamo: Amazon's highly available key-value store. ACM SIGOPS operating systems review, 41(6), 205-220. doi.org/10.1145/1323293.1294281

Demchenko, Y., De Laat, C., & Membrey, P. (2014, May). Defining architecture components of the big data Ecosystem. In 2014 International conference on Collaboration Technologies and Systems (CTS) (pp. 104-112). IEEE. doi.org/10.1109/CTS.2014.6867550

Doug, L. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety, META group.

Elgendy, N., & Elragal, A. (2014, July). Big data analytics: A literature review paper. In Industrial conference on data mining (pp. 214-227). Springer, cham. doi.org/10.1007/978-3-319-08976-8_16

Emani, C. K., Cullot, N., & Nicolle, C. (2015). Understandable big data: A survey. Computer science review, 17, 70-81. doi.org/10.1016/j.cosrev.2015.05.002

Franks, B. (2012). Taming the big data tidal wave: Finding opportunities in huge data streams with advanced analytics. John Wiley and Sons, https://onlinelibrary.wiley.com/doi/book/10.1002/9781119204275.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods and analytics. International journal of information management, 35(2), 137-144. doi.org/10.1016/j.ijinfomgt.2014.10.007

Gantz, J., Reinsel, D., Chute, C., Schlichting, W., McArthur, J., Minton, S., Xheneti, I., Toncheva, A. and Manfrediz, A. (2007) The Expanding Digital Universe: A Forecast of Worldwide Information Growth through 2010. Technical Report, 12, 634-638. https://www.tobb.org.tr/BilgiHizmetleri/Documents/Raporlar/Expanding_Digital_Universe_IDC_WhitePaper_022507.pdf

Geng, B., Li, Y., Tao, D., Wang, M., Zha, Z. J., & Xu, C. (2011). Parallel lasso for large-scale video concept detection. IEEE transactions on multimedia, 14(1), 55-65. doi.org/10.1109/TMM.2011.2174781

Groves, P., Kayyali, B., Knott, D., Kuiken, S. V. (2013). The 'big data' revolution in healthcare, Center for US Health System Reform Business Technology Office.

Khan, S., Ilyas, Q. M., & Anwar, W. (2009, December). Contextual advertising using keyword extraction through collocation. In Proceedings of the 7th international conference on frontiers of information technology (pp. 1-5). doi.org/10.1145/1838002.1838081

Mishra, S., & Tyagi, A.K. (2022) The Role of Machine Learning Techniques in Internet of Things-Based Cloud Applications. In: Pal S., De D., Buyya R. (eds) Artificial Intelligence-based Internet of Things Systems. Internet of Things (Technology, Communications and Computing). Springer, Cham. doi.org/10.1007/978-3-030-87059-1_4

L'heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. (2017). Machine learning with big data: Challenges and approaches. Ieee Access, 5, 7776-7797. doi.org/10.1109/ACCESS.2017.2696365.

Li, X., & Yao, X. (2011). Cooperatively coevolving particle swarms for large scale optimization. IEEE Transactions on Evolutionary Computation, 16(2), 210-224. doi.org/10.1109/TEVC.2011.2112662

Lin, F., & Cohen, W. W. (2010, January). Power iteration clustering. In ICML.

Lesk, M. (1997). How much information? Report on American consumers, Global Information Industry Center, University of California, San Diego.

Mansouri, N., & Fox, G. C. (1991). Parallel genetic algorithms with application to load balancing for parallel computing. https://surface.syr.edu/eecs_techreports/128/

Mariscal, G., Marban, O., & Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review, 25(2), 137-166. doi:10.1017/S0269888910000032.

Nambiar, R., Bhardwaj, R., Sethi, A., & Vargheese, R. (2013, October). A look at challenges and opportunities of big data analytics in healthcare. In 2013 IEEE international conference on big data (pp. 17-22). IEEE. doi.org/10.1109/BigData.2013.6691753

Neumeyer, L., Robbins, B., Nair, A., & Kesari, A. (2010, December). S4: Distributed stream computing platform. In 2010 IEEE International Conference on Data Mining Workshops (pp. 170-177). IEEE. doi.org/10.1109/ICDMW.2010.172

Oussous, A., Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2018). big data technologies: A survey. Journal of King Saud University-Computer and Information Sciences, 30(4), 431-448. doi.org/10.1016/j.jksuci.2017.06.001

Oyelade, J., Isewon, I., Oladipupo, F., Aromolaran, O., Uwoghiren, E., Ameh, F., ... & Adebiyi, E. (2016). Clustering algorithms: Their application to gene expression data. Bioinformatics and Biology insights, 10, BBI-S38316. doi.org/10.4137/BBI.S38316.

Pramod, A., Naicker, H. S., & Tyagi, A. K. (2021). Machine learning and deep learning: Open issues and future research directions for the next 10 years. Computational analysis and deep learning for medical care: Principles, methods and applications, 463-490. doi.org/10.1002/9781119785750.ch18

Quackenbush, J. (2002). Microarray data normalization and transformation. Nature genetics, 32(4), 496-501. doi.org/10.1038/ng1032

Bohn, R. E., & Short, J. E. (2009). How much information? Report on American consumers, Global Information Industry Center, University of California, San Diego.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. Health information science and systems, 2(1), 1-10. doi.org/10.1186/2047-2501-2-3

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23(4), 3-13. https://cs.brown.edu/courses/cs227/archives/2017/papers/data-cleaning-IEEE.pdf#page=5

Rao, B. P., Saluia, P., Sharma, N., Mittal, A., & Sharma, S. V. (2012, December). Cloud computing for Internet of Things & sensing based applications. In 2012 Sixth International Conference on Sensing Technology (ICST) (pp. 374-380). IEEE. doi.org/10.1109/ICSensT.2012.6461705

Taylor, R. C. (2010). An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. BMC bioinformatics, 11(12), 1-6. doi.org/10.1186/1471-2105-11-S12-S1

TAFFBDC. (2012). Demystifying bigdata: A practical guide to transforming the business of Government. https://bigdatawg.nist.gov/_uploadfiles/M0068_v1_3903747095.pdf

Tsai, C. W., Lai, C. F., Chao, H. C., & Vasilakos, A. V. (2015). Big data analytics: A survey. Journal of big Data, 2(1), 1-32. doi.org/10.1186/s40537-015-0030-3

Tyagi, A. K. (2019, February). Machine learning with big data. In Machine Learning with big data (March 20, 2019). Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur-India. doi.org/10.2139/ssrn.3356269

Tyagi, A. K., & Chahal, P. (2020). Artificial intelligence and machine learning algorithms. In Challenges and applications for implementing machine learning in computer vision (pp. 188-219). IGI Global. doi.org/10.4018/978-1-7998-0182-5.ch008

Xu, G., Zhang, Y., & Li, L. (2011). Web content mining. In Web Mining and Social Networking (pp. 71-87). Springer, Boston, MA. doi.org/10.1007/978-1-4419-7735-9_4

Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., & Vasilakos, A. V. (2016). Big data: From beginning to future. International Journal of Information Management, 36(6), 1231-1247. doi.org/10.1016/j.ijinfomgt.2016.07.009

Appendix A

Table 1: Various categories of big data

Classification	Description
Datastores	
Document-oriented data	Archive situated information stores are principally intended to aggregate and recover a set of information or data. It underpins complex information structures in various standard configurations, as JSON, XML and furthermore, parallel structures (e.g., PDF and MS Word). A document-orient information store is like a tuple or column in a social data set. But principle advantage is that it is more adaptable and can get archives dependent on their substance (e.g., Mongo DB, Simple DB and Couch DB)
Column-oriented database	A section arranged information base stores its substance in segments instead of lines. Here trait esteems holding a similar segment are put away conterminously. Segment arranged is not the same as customary data set frameworks which store whole records in a steady progression (Abadi *et al*., 2009), for example, big table (Chang *et al*., 2008)
Graph database	A graph database is intended to store and address information that uses a chart model which comprises vertices, edges, and properties identified with one another utilizing relations (Neubauer, 2010). (e.g., Neo4j)
Key-value	Key-value is an option relational information base framework that stores and gets to information intended to scale to an enormous size (Seeger and Ultra-Large-Sites, 2009). One of the significant models for this stockpiling framework is Dynamo (DeCandia *et al*., 2007). Some, of the amazon administrations, utilize this framework. Likewise, (Das *et al*., 2010) proposed an adaptable key-esteem store to help value-based multi-key access utilizing solitary key access upheld by key-value for use in G-store plans. In (Lin and Cohen, 2010), creators introduced a versatile bunching technique to play out an enormous assignment in datasets. Different instances of key-value stores are Apache Hbase (Taylor, 2010), Apache Cassandra (Lakshman and Malik, 2011) and Voldemort. Hbase utilizes HDFS, an open-source variant of Google's Big Table based on Cassandra. Hbase stores information in tables, columns, and cells. Lines are arranged by line key and every cell in a table is indicated by a line key, a section key, and a variant, with the substance contained as an un-deciphered exhibit of bytes

Table 1: Continue

Data sources	
Social media	Web-based media can be treated as the source of data which assists with trading data and thoughts in practically associated networks, as synergistic tasks, websites and microblogs, Facebook and Twitter
Machine data	Machine information is considered as data naturally created from the machine (either equipment or programming) like PCs, sensors, or different gadgets, without human mediation
Sensing	The main uses of sensing devices are measuring physical quantities and converting them into signals.
Transaction data	Transaction data consist of time-oriented data, such as financial and work data
Internet of Things (IoT)	IoT implies recognizable proof of gadgets as a piece of the Internet. These gadgets can be cell phones, advanced cameras, tablets, and so on at the point when these gadgets associate with each other over the Internet, they can do smart cycles and administrations which uphold fundamental, monetary, ecological, and wellbeing needs Countless gadgets associated with the Internet give a huge measure of administration without human intercession and furthermore, deliver immense measures of information and data (Rao *et al*., 2012)
Content format	
Structured	information is regularly overseen by social data sets like SQL for example organized question language. Principle utilization of this language is overseeing and questioning information in RDBMS structured information is not difficult to gather, inquire about, store, and inspect. Instances of structured information tabulated
Semi-structured	Semi-structured information is information that doesn't follow the conventional data set framework. They might be as organized information which is not detailed in social data set models, like tables. Semi-structured information assortment measure with the end goal of the investigation isn't equivalent to that of a fixed record designs. It utilizes complex principles that powerfully choose the following interaction after catching the information (Franks, 2012)
Unstructured data	Unstructured data do not follow a particular format, It includes text messages, videos, and social media data. Since the use of smartphones is more, the unstructured type of data is also generated in huge quantities Analysis and interpretation of such data are difficult
Data staging	
Cleaning	is a data pre-processing technique for removing unwanted data like noise (Yaqoob *et al*., 2016; Rahm and Do, 2000)
Transform	Transformation is the process by which data is converted into a suitable format for analysis (Yaqoob *et al*., 2016).
Normalization	is a pre-processing method for scaling the data to minimize redundancy (Quackenbush, 2002)
Data processing	
Batch	It processes the task as batches without human intervention. MapReduce-based systems have been implemented with batch jobs (Chen *et al*., 2012). It helps to process a huge volume of data
Real-time	It measures the information quickly when the information is inputted. Quite possibly the most well-known and amazing ongoing interaction-based enormous information apparatuses are straightforward adaptable streaming frameworks (S4) (Neumeyer *et al*., 2010). S4 is a dispersed figuring stage that permits software engineers to helpfully create applications for preparing ceaseless unbounded surges of information. S4 is a versatile, mostly deficient lenient, universally useful, and pluggable stage

Table 2: Addressing the big data challenges with emerging methodologies, technologies, and solutions.

Technology/big data challenge	Storage	Transfer	Management	Preprocessing/ processing	Analysis	Visualization	Integration	Architecture	Security/privacy	Quality	Efficiency Cost/energy
Distributed file/storage system	Y		Y	N							
Cloud monitoring and tracking									N		Y
Near field communication		Y									
NoSQL database systems	Y		Y								
Search, query, indexing and data model design	Y	Y	Y	Y	Y						
Map Reduce (Hadoop) system				Y	Y	Y	Y	Y			
Parallel programming languages				Y	Y	Y	Y				
big data analytics and visualization			N		Y	Y	Y				
Semantics			Y		Y				Y	Y	
Mobile data and computing		Y		Y	Y						Y
Internet of things		Y		Y					N		
Computing infrastructure			Y	Y	Y	Y	Y				Y
Cloud computing	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y
Management and processing architecture	Y		Y	Y	Y	Y	Y	Y			
Remote collaboration	Y	Y	Y	Y	Y	Y	Y	Y	Y		
Anything as a service	Y	Y	Y	Y	Y	Y	Y	Y			
Resource auto-provision, scaling and scheduling	Y		Y	Y	Y	Y	Y	Y			Y
Spatiotemporal optimization	Y	Y	Y	Y	Y	Y	Y	Y			Y
Statistical analyses, machine learning and data mining			Y								