Journal of Big Data

SURVEY

Open Access

A brief survey on big data: technologies, terminologies and data-intensive applications

Hemn Barzan Abdalla*

*Correspondence:
habdalla@kean.edu

Wenzhou-Kean University, Wenzhou, China

Abstract

The technical advancements and the availability of massive amounts of data on the Internet draw huge attention from researchers in the areas of decision-making, data sciences, business applications, and government. These massive quantities of data, known as big data, have many benefits and applications for researchers. However, the use of big data consumes a lot of time and imposes enormous computational complexity. This survey describes the significance of big data and its taxonomy and details the basic terminologies used in big data. It also discusses the technologies used in big data applications as well as their various complexities and challenges. The survey focuses on the various techniques presented in the literature to restrain the issues associated with big data. In particular, the review concentrates on big data techniques in accordance with processing, security, and storage. It also discusses the various parameters associated with big data, such as availability and velocity. The study analyses big data terminologies and techniques in accordance with several factors, such as year of publication, performance metrics, achievement of the existing models, and methods utilized. Finally, this review article describes the future direction of research and highlights big data possibilities and solicitations with a detailed sketch of the big data processing frameworks.

Keywords: Internet of Things (IoT), Big data, Application, Challenges, Techniques

Introduction

Big data has become the latest and eminent research topic because of its widespread application and use across various domains. According to the report presented by Gartner in 2013, big data holds a prominent position among innovative technologies and has been listed among the leading trend technologies from 2013 to 2018" [88]. Big data is characterized as an assortment of enormous databases, which presents difficulties when determining the best class of information for handling deadlines or customary information-preparing stages. In 2012, Gartner gave a more point-by-point definition of big data: "Big data is described as high-speed, or potentially high-volume, high-assortment data resources that demand new types of handling measures to empower improved dynamic, knowledge-revelation and cycle advancement." For the most part, an informational collection that can effectively perform catch, corporation, examination, and perception at the current advancements can be called "big data" [2]. As per

the International Data Corporation (IDC) [1], the big data innovation market generated about $32.4 billion in 2017. Today, with the rapid development of enormous information science and innovations, various information mining techniques, AI-based calculations, and open-source stages and apparatuses, big dataset advances have been created and made accessible to those who need them. The accessibility of these big datasets, or "big data", for use in big data applications, which recommend effective information processing and application administration, has enormously expanded the scope of meeting business necessities and requests in individuals' everyday lives [3].

In the modern world, big data is utilized in almost all the application frameworks, such as suggestion frameworks, forecasts, perceived patterns, and factual report applications. Emerging enormous data processing and administrations can be utilized in various domains and applications, including organizational management, media center, environmental condition, educational organization, biomedical services, medical services and life sciences, online media and systems administration, smart city transportation, and data transferring [2, 7]. Moreover, enormous information-based applications are generally utilized as a suggestion, a prediction [9], or a choice framework [10]. Thus, the expression "information-driven dynamic" emerges to portray the strategies for gathering and evaluating information to direct or further enhance choices [9]. The dynamic decision-making framework helps marketers focus on track-advertising, protection suppliers focus on offering customized protections to their clients, and healthcare providers focus on providing patients with high-quality, low-cost therapy [7, 8]. This includes the evaluation of non-conditional and undetermined information, such as product descriptions or customer audits. For instance, information experts investigate the vast information gathered from web-based media to test a specific theory.

As a result, one can decide these hypotheses' worth, legitimacy, and attainability and set up the designs for executing them [11]. Although there have been advancements in information storing, assortment, evaluation, and calculations related to foreseeing human conduct, it is crucial to understand the hidden driving forces and the managing factors, such as social norms, framework, law, market-accepted practices, and engineering, that aid in creating powerful models that can deal with extensive information and, thus, boost the forecast accuracy [13]. However, because of the tremendous volume of produced information, the quick speed at which this information becomes available, and the enormous assortment of heterogeneous information, the large information-based applications bring new difficulties and issues for Quality Assurance (QA) engineers.

For instance, due to the enormous scope of information size and the element of practicality, it is a challenging task to accept the correctness of a major information-based expectation framework. Along these lines, big information quality approvals and large information-based application framework quality affirmations become significant concerns that require further examination. Even though several previous papers have studied information quality and information quality affirmation [15–20], very few articles center around the approval for enormous information application quality. There is a rising need in research work for quality investigation issues and quality confirmation answers for large information applications [1, 31, 42]. The subsequent significant topic of research in this field is the programmed age of the metadata, which portrays the recorded information, and describes how it is recorded and estimated. Security [3,

21, 23] is a significant concern, particularly with regard to big data. For instance, there are now laws to protect the privacy of patients in many regions. There is an increasing apprehension of improper utilization of individual information, especially when it is combined from various sources. In addition, obtaining big data has its unique difficulties, which are quite different from those related to conventional information [2]. The review of the big data terminologies, associated technologies, and applications is enumerated in this article.

This research article describes the big data techniques concerning storage, processing, and security. Moreover, the article discusses various parameters associated with big data, such as availability and velocity. The study analyses big data terminologies and techniques under several factors, such as year of publication, performance metrics, achievement of the existing models, and methods utilized. Finally, this review paper describes the future direction of research, highlighting big data opportunities and applications with a detailed sketch of the big data processing frameworks.

The organization of the review paper is as follows: "Introduction" Section defines big data and its different aspects. "The review paper has several objectives" Section presents an analysis based on big data tools and Hadoop architecture. "Analysis based on big data tools and Hadoop architecture" Section describes the review-based big data application in various domains. "Review and analysis of big data technologies" Section highlights the challenges associated with big data technologies. "Analysis and discussion" Section describes big data analytics issues and challenges, while "Challenges associated with big data applications" Section provides a comparative discussion with existing articles. Finally, "Big data analytics: issues and challenges" Section concludes the articles.

The review paper has several objectives:

- To review the techniques of big data processing application domains based on machine learning, deep learning, cloud computing, fog computing, edge computing, concentric computing, Internet of Things (IoT), big data analytics, and Hadoop.
- To analyze big data using tools such as NoSQL, Cassandra, Hadoop, Strom, Spark, Hive, and OpenRefine.
- To evaluate big data in terms of performance analysis based on accuracy, resource utilization, RMSE, SD, scalability, speed, security, processing time, TPR, FPR, throughput, detection rate, energy consumption, mean, item loading, delay, stability, and SINR.

Big data definition

In contrast to conventional data, the term "big data" alludes to developing enormous informational indexes that incorporate heterogeneous configurations, i.e., organized, unorganized, and semi-organized information. Extensive data has a mind-boggling nature that requires incredible advances and progressed calculations. Therefore, the conventional consistent business Intelligence (BI) technology can not be used for big data applications. Big data is produced from different origins and in numerous configurations (e.g., recordings, reports, remarks, and logs). Huge data collections comprise

information that is organized and unorganized, private and public, neighborhood and far off, divided and secret, and complete and inadequate [4, 18]. Most researchers and specialists characterize big data by five accompanying principle attributes, called the 5 V's: volume, velocity, variety, veracity, and value. The five V's in big data are briefly described below.

Volume

Volume represents the greatness of information. Large information sizes account for different terabytes and petabytes. A review directed by the international organization International Business Machines Corporation (IBM) in 2012 disclosed that over a portion of the 1144 respondents considered datasets of more than one terabyte to be colossal information. One terabyte stores as much information as would enclose on 220 DVDs or 1500 CDs, sufficient to reserve up to 16 million Facebook images, and one petabyte is equivalent to 1024 terabytes. The report states that Facebook creates billions of images per second, equivalent to one petabyte. A prior estimation states that Facebook stores more than 260 billion images, utilizing space of about 20 petabytes. The description of big data volumes is contingent and differs depending on factors such as how the data is sorted and the time of information. It may become difficult to maintain the enormous data in the future because the capacity limits may be incremented, permitting significantly greater informational indexes to be captured. Furthermore, the kind of information discussed under assortment characterizes what is implied by "large." Two datasets of a similar size might require different information, and the executive innovations rely on how the databases are sorted. In addition, the meaning of enormous information depends on the business. Therefore, these factors make it impractical to define a particular limit for large information volumes [8].

Variety

Variety is defined as the constructional diversification in a dataset. Modern technical advancements permit firms to utilize different types of organized, unorganized, and unstructured information. The organized information, which comprises 5% of every current datum, alludes to the plain information found in social datasets or spreadsheets. Text, pictures, sound, and video are instances of unstructured information, which occasionally lack the primary associations needed by machines for investigation. The organization of semi-organized information doesn't adjust to severe guidelines, bisecting the interconnectedness between progressively unorganized and organized information. The eXtensible Markup Language (XML), a text-based language for trading information on the Web, is a normal illustration of semi-organized information. XML records contain client-characterized information labels, which make them machine-meaningful. They have an undeniable degree of assortment, a characterization normal for enormous information. Associations have been storing unstructured information from inside sources (e.g., sensor information) and external sources (e.g., online media) for a long time. Nonetheless, the modern creative viewpoint encourages the development of new information, executive advances, and examination, which empower associations to use information in their business measures. For example, facial acknowledgment innovations enable physical retailers to obtain knowledge about store congestion, the sex or age of their clients,

and their in-store development designs. This vital data is utilized in item advancements, situations, and staffing choices. The clickstream information gives an abundance of data and examples of client conduct to online retailers, which educates them on the circumstance and grouping of pages viewed by a client. Utilizing massive information investigation, even small and medium-sized endeavors (SMEs) can mine monstrous volumes of semi-organized information to develop web architectures further and implement viable strategic pitching and customized item suggestion frameworks.

Velocity

Velocity is defined as the rate at which information is created and the speed at which it should be evaluated and followed upon. The extension of advanced equipment, for example, personal digital assistants (PDAs) and sensors, has prompted a phenomenal pace of data production and is driving a growing need for continual research and proof-based organization. Even traditional retailers are creating high-recurrence information. For instance, Wal-Mart measures more than 1,000,000 exchanges per hour. The information radiating from cell phones and moving through versatile applications produces streams of data that can be utilized to create ongoing, customized offers for regular clients. This information gives detailed data about clients, such as geospatial area, socio-economics, and past purchasing behaviors, which can be investigated continuously to raise genuine client esteem [10].

Veracity

Veracity indicates the accuracy and the reliability of the data, which is collected from various sources. There are some possibilities of mixing up the inaccurate and unprocessed data. Hence, data veracity indicates the reliability and uncertainty level of the data [7].

Value

The value represents the defining attribute of big data. Evaluation of large volumes of data provides a high value, which in turn increases the knowledge obtained from the data [10].

Visual analytics in big data

Visual analytics (VA) concentrates on assisting the interaction and exploration in the analysis of big complex data. Visual Analytics of Big data depends on three main layers: the Visualization layer, the analytics layer, and the data management layer. A brief description of each layer is given in the following subsection.

Visualization

The term visualization is defined as the utilization of the collective visual representation of the conceptual data to increase the perception. The functionality and the aesthetic form are required for the easy conveyance of the information. The information such as variables and attributes is abstracted from the data in semantic form. One of the online marketing platforms known as eBay utilizes the data visualization tool to understand all the data generated by their customers. Hence, the employees in eBay use the

visualization tool to monitor the recent customer feedback to provide a better quality of services. Conducting visualization for Big data is a complex process due to its high dimensionality and large size. Visualization needs consideration as it needs to manage the dataset with high volume and speed [87, 89].

Analytics

Analytics is defined as the process of obtaining conclusions from big data through data evaluation. Researchers adopt several processing methods in the evaluation process to attain the preferable result. If the data is small, the review can be performed quickly, and the data is visualized using the charting facility. Hence, the integration of the analytical environment with the visual environment is widely utilized in industries and research. If the data is massive, the integration of visualization and analytics may not perform well and could generate scalability issues [89].

Data management

Data management is a crucial aspect of VA applications as it helps to manage the life cycle of data. Data management contributes to the quality assurance, data retrieval, and preservation of data over time. Conventional data management tools fail to handle big data, which is complex and large [89].

"The review paper has several objectives" Section provides a brief explanation of the existing tools for processing big data, along with their benefits and drawbacks.

Analysis based on big data tools and Hadoop architecture

The first part of this section provides a brief description of the Big data tools, and the second part of the section explains the technologies in the big data based on the apache tool. The third part of the section explains the Hadoop Ecosystem in Big data. The tools and technologies with their pros and cons are analyzed in this part, which helps the researchers to select the better tools for processing the big data.

Big data tools

The continued advancement of business activities excessively depends on the intensive evaluation of big data. The analysis of big data assumes a significant part in the decision-making process to enhance the advancement and prosperity of the association [9]. In any case, sheer computing measures of information with the guide of customary information computing devices fail to provide productive results, and the device is not attainable. Therefore, several big data tools have been developed in the past years that assist associations and data researchers in inferring the information-driven choices productively and cost-effectively. An assortment of big data analytics [4–7, 13, 25–41] apparatuses are utilized by specialists to work with information storage, information management, information purging, information mining, data forecasting, and data endorsement. This section briefly describes the tools utilized to analyze BD [33, 50].

NoSQL

Generally, the SQL (structured query language) is extensively utilized to restrain and analyze structured data. However, the immense advancement of indefinite data has

originated the rise of unstructured information-analytical apparatuses. Eventually, SQL (NoSQL) emerged to deal with unorganized data models productively. NoSQL information bases do not mainly stick to architecture while storing indefinite data. Subsequently, the segment values of the table shift are by every information record (line). Because of this architecture-less nature, the system balances the consistency over-momentum, adaptation to internal failure, and accessibility. Although NoSQL has acquired huge prevalence during the last few years, challenges emerge from low-level inquiry dialects, and the absence of normalized interfaces is yet to be considered [12].

Cassandra

Cassandra is the type of NoSQL, non-proprietary, and disseminated dataset that manages the exceptionally enormous datasets across various providers. The magnificence of Cassandra is regarding the low failure, guaranteeing high accessibility under any conditions. Thus, the group of experts prefers Cassandra when versatility and accessibility highlights are pivotal without affecting the execution performance. In addition, Cassandra enables information replication across various clouds [19] or server farms to guarantee lower dormancy and adaptation to internal failure.

Hadoop

Hadoop is a system that comprises an assortment of programming libraries, which integrates the different programming models to enable the disseminated computation of enormous datasets. Versatility is the significant advantage associated with the Hadoop system, where the dispersed repository is independently called the Hadoop Distributed File System (HDFS). This Hadoop framework is a highly accessible analytical tool independent of the hardware equipment. Instead, it consolidates programming libraries to distinguish, recognize, and restrain the insufficiency at the implementation layer. The Hadoop system comprises the normal libraries, stockpiling libraries, Hadoop YARN, and Hadoop MapReduce to compute the vast datasets.

Strom

The storm is a continuous free non-proprietary dispersed processing framework which controls the enormous measure of streaming information that resembles the cluster computation of Hadoop. The magnificence of the storm is handled using the programming language, and for guaranteeing legitimate ongoing investigation experience, the storm coordinates the current queuing components in addition to the dataset innovations. The streaming procedure of storms requires intricate and subjective packets at every algorithmic stage. Real-time investigation, ceaseless execution, and online (Machine Learning) ML are a portion of the key administrations recommended by storm.

Spark

The spark is a widely accepted open-source dispersed, cluster-processing system. Spark ensures better production in both stream information computing and cluster computing. Spark coordinates with GraphX, SQL, Spark streaming segments, and MLib. Spark validates the information from different data sources and executes on different stages,

like Mesos, Yarn, and Hadoop. Since the same code regulation may be applied to both constant computing and cluster computing, the spark is considered productive in the storm evaluation. Notwithstanding, the storm has acquired prevalence in terms of inactivity with fewer limitations.

Hive

Hive is a cross-stage information distribution programming center that works on Hadoop, which empowers information evaluation and information querying in numerous storage areas and document frameworks coordinated with Hadoop. To inquire over conveyed document frameworks, Hive gives SQL reflection through HiveQL, a SQL-like inquiry language. Subsequently, solicitations do not require processing on the questions, utilizing lower-grade Java APIs. HiveQL straightforwardly changes over questions into Apache Tez, MapReduce, and Spark. Besides, Hive offers lists to develop the inquiry computing speed further.

OpenRefine

The OpenRefine is the non-proprietary independent solicitation recently recognized as Google Refine. The framework is broadly utilized in the present information-operated world to scrub noisy and crude information and change information starting with one structure and then the next. OpenRefine is related to the conventional Relational Database Management Systems (RDBMS) tables, which store lines of the information under sections. Nonetheless, it digresses from the traditional situation, as no equations stockpile the cells under sections. Moreover, the equations utilized in the OpenRefine change the information, and the changes are permitted only once [6].

Version of apache tools

This section enumerates different technologies employed in big data based on the apache tool in detail and lets us take a deep insight into the technologies.

Apache flume

Apache Flume [20, 43, 44] is a dispersed, dependable, and accessible framework for effectively gathering, comprehending, and transferring a lot of log information from a wide range of sources to the incorporated information store. The main advantage of apache flume is that it is false tolerant, reliable, scalable, and accessible for different sinks and sources. Furthermore, the Apache Flume helps to store data in centralized stores such as HDFS and HBase [10, 47]. The main limitation is that there are no ordering guarantees for any of the events. Hence, in the future, research has to overcome the zero ordering guarantees to ensure the proper delivery of the data packets.

Apache sqoop

Apache Sqoop is a command-line interface (CLI) device intended to move information among Hadoop and social datasets. Sqoop can use implicit information from the RDBMS [17], such as Oracle Dataset or MySQL stockpiled into HDFS, and then disseminate the information back once MapReduce has altered the information. The Sqoop additionally brings information into Hive and HBase and is associated with an RDBMS

through the Java database connectivity (JDBC) connector and depends on the RDBMS to portray the dataset architecture for information to be imported. The Apache Sqoop allows faster, effective, and low-cost offload processes such as ETL (Extract, Load, and Transform). In the future, more consideration must be given to mitigating import and export failure, which degrades the performance of the Apache Sqoop system.

Apache pig

Apache's Pig is a significant venture, which inherits the top position of Hadoop, and permits more elevated level programming language to utilize the library of Hadoop's MapReduce. Pig gives the prearranging language to portray tasks, such as perusing, separating and changing, consolidating, and composing information, which precisely executes the similar activities that MapReduce was initially intended-for. Rather than communicating these tasks in a vast number of lines of Java code, MapReduce is utilized straightforwardly, and Apache Pig allows the clients to communicate in a different language that differs from the normal programs, such as Perl or slam script. Unlike other SQL, Pig does not require data architecture, so it is appropriate to restrain the unstructured information. Yet, Pig can, in any case, use the worth of architecture, assuming that the clients need to contribute it. The PigLatin is comparatively accomplished, like SQL, and the turning culmination requires restrictive formulation, unlimited memory frameworks [18], and encompassing framework. The lack of an integrated development environment (IDE) is the main drawback of the Apache pig system. The future research in the Apache pig is to develop the IDE, which offers functionalities to compile the pig scripts.

Apache hive

Hive is an innovation created by Facebook that transforms Hadoop into an information distribution center that terminates with an ascent of the SQL for inquiring. HIVEQL is the decisive language in SQL. In the PigLatin, the clients will determine the information stream. However, in Hive, the clients portray the outcome required for the client, and the hive sorts out some way to assemble an information stream to accomplish the required outcome. Unlike Pig, Hive requires construction, yet the clients are not restricted to just a single composition. Like SQL and PigLatin, HiveQL itself is a comparatively terminated language. The high latency is the main drawback of the Apache Hive, which should be considered in future research.

Apache ZooKeeper

Apache Zoo Keeper is an attempt to create and sustain open-source providers, which empowers profoundly dependable disseminated coordination. It enables the disseminated design administration, a synchronization administration, and a naming vault for conveyed frameworks. The disseminated applications utilize the ZooKeeper to store and intercede updates to implicit the data. The Zoo Keeper is particularly quick with responsibilities, where peruses of the information are more normal than composed. The best read/compose proportion is found to be 10:1, and the Zoo Keeper is recreated over the group of hosts known as a troupe, and the providers are aware of one another, and there are no failure nodes. There is a chance for accidental data loss as it fails to support the

redundant number of pods. Hence, the future research direction is to develop a platform to support the redundant number of pods to avoid accidental data loss.

Apache Cassandra

The Apache Cassandra is the other non-proprietary source NoSQL data set arrangement that has acquired modern innovations, which can deal with enormous information necessities. It is an extremely versatile and elite information base administration framework that can deal with enormous ongoing applications that drive key frameworks for current and effective organizations. It has established a for-scale design that can deal with petabytes of data and millions of clients/activities each second as effectively as maintaining a modest measure of information and client traffic. Apache Cassandra is additionally outfitted with an adaptable/powerful blueprint plan that obliges all arrangements of large information applications, including organized, semi-organized, and unstructured information. Cassandra addresses information through dynamic segment families that oblige all changes on the web. The latency of Apache Cassandra due to excessive requests and data tends to be minimized in future research directions.

Apache Hadoop

The Apache Hadoop processing media center is a structure that empowers the propagated handling of enormous informational indexes across groups of PCs. It is intended to increase from single workers to a huge amount of machines, with each associated neighborhood estimation and proportions. The essential idea is to permit a solitary question to discover and gather results from all the cluster individuals. This model is precisely appropriate for Google's framework of search support. One of the most serious mechanical difficulties in programming frameworks research today is giving systems the capacity to control and data recovery on an enormous amount of information. MapReduce is the hugely adaptable, parallel computing structure generally utilized with Hadoop and different parts, for example, the YARN and HDFS. YARN can be depicted as a huge scope, disseminated working framework for enormous information executions. With the development of Hadoop, the cluster arranged, disk-concentrated MapReduce impediments have attained more clear idea as big data examination moves to all the more ontime application, stream handling, and progressed executions. MapReduce enables the quickest, most cost-efficient, and most versatile component for bringing results back. Today, the vast majority of the main innovations for overseeing "large information" are created on MapReduce. MapReduce has few adaptability restrictions. However, its straightforward utilization requires composing and keeping a great deal of code. Hence, the future research direction is to reduce the number of codes to minimize the computational delay.

Apache Splunk

Apache Splunk is a widely utilized tool to explore, investigate, and detail the timeseries text information derived from machine information. Splunk programming is used to address at least one central IT capacity: application executives, security, consistency, IT tasks control, and exploring business inquiry. The Splunk motor has been enhanced for rapidly organizing and maintaining unstructured information stacked

into the framework. In particular, Splunk utilizes an insignificant construction for continued information, and the occasions only include the crude occasion text, suggested timestamp, source (the filename for document-based data sources), source type (a sign of the overall kind of information), and host (where the information started). The Apache framework helps to build the real-time data application. When information enters the Splunk framework, it rapidly continues through computing, is persevered in its crude structure, and is filed by the above fields alongside every one of the keywords in the crude occasion text. Scheduling is a fundamental component of the standard "super-grep" use case for Splunk, but it also speeds up most restoring processes. On these crude occasions, any more modern computing is permitted till the end of the search. This serves four significant objectives: execution speed is expanded as negligible preparation is performed, carrying new information into the framework is a generally low exertion practice as no pattern arranging is required, the first information perseveres for simple examination, and the framework is versatile to change as information parsing does not require reloading or re-ordering the information [2]. In the future, monitoring tools should be enhanced to support a large volume of data.

Big data and Hadoop ecosystem

Hadoop capabilities

The Hadoop framework is a notable big data innovation that significantly supports local areas. It has been designed to avoid the low exhibition and intricacy experienced when limiting and dissecting big data using conventional innovations. One primary benefit of Hadoop is its ability to quickly deal with huge informational collections because of its equal groups and conveyed document framework. Not normal for customary advancements, Hadoop does not replicate the entire far-off information to execute calculations in memory. Rather, Hadoop executes errands, where information is stored. Hence, Hadoop mitigates the correspondence load of organizations and providers. Another benefit of Hadoop is its capacity to run programs while guaranteeing adaptation to non-critical failure, generally experienced in dispersed climates. It forestalls information loss by imitating workers' information to ensure that. The force of the Hadoop stage depends on two fundamental sub-segments: the Hadoop Distributed File System (HDFS) along with MapReduce structure (clarified in the accompanying segments). Similarly, clients can add modules on top of Hadoop based on the situation indicated by their targets and their application prerequisites (e.g., limits, exhibitions, dependability, adaptability, and security). Generally, the Hadoop organization has added to enhance its biological system with a few non-proprietary modules. Similarly, IT merchants give extraordinary endeavors solidifying highlights conveyed inside Hadoop appropriations [86].

Information storage layer Hadoop depends on the framework, such as HBase and HDFS, to stockpile information, its document framework HDFS, and a self-subsistent data set called Apache HBase.

(a) HDFS

The HDFS is an information-stockpiling framework that upholds up to numerous nodes in a bunch and gives a practical and dependable stockpiling capacity. It can deal with both organized and unstructured information and hold tremendous volumes (i.e., stockpiled documents can be greater than a terabyte). Nonetheless, clients should know that HDFS does not establish a general-purpose framework because it was designed for high-idleness activities cluster computing. Furthermore, it does not give quick record queries in documents. The main advantage of HDFS is its versatility across heterogeneous equipment and programming stages. Furthermore, HDFS assists with diminishing organization clog and increment framework execution by moving calculations close to information stockpiling. It guarantees additional information replication for adaptation to internal failure. Those highlights clarify its wide reception. HDFS depends on master–slave design. It allocates huge amounts of information to the group. Indeed, the group has a novel expert (NameNode) that controls document framework activities and many slaves (DataNodes) that oversee and arrange information stockpiling on individual register nodes. To give information accessibility, Hadoop is located on information replication.

(b) HBase

HBase is an appropriated independent dataset and a non-proprietary project developed based on the performance of HDFS, designed for low-delay activities. Hbase depends on the section arranged key/esteem information model. It can uphold high table-streamline rates and augment on a level plane in conveyed groups. HBase gives an adaptable, organized facilitating to enormous tables in a BigTable-like organization. Tables store information coherently in lines and sections. The advantage of such tables is that they can deal with billions of lines and a huge number of segments. With the help of HBase, multiple ascribes may be assembled into section families, allowing a segment family's components to be stored together. This methodology is not quite the same as a line situated social dataset, where all sections of a line are stored together. Hence, HBase is more adaptable than social information bases. All things being equal, HBase has the advantage of permitting clients to acquaint refreshes with better handle changing applications' prerequisites. Each table should have a characterized pattern with a Primary Key that is utilized to get to the Table. The line is recognized by the table's name and the start key, while sections might have a few adaptations for a similar line key. Hbase gives many highlights, such as ongoing inquiries, normal language search, predictable admittance to Big Data sources, straight and measured adaptability, and programmed and configurable fragment of tables. It is known for some Big Data arrangements and information-driven sites, like Facebook's Informing Platform. HBase incorporates Zookeeper to coordinate administrations and execute as the Zookeeper occasion naturally. In addition to HBase, HDFS has a MasterNode that manages the bunch and slaves that reserve portions of the tables and perform the procedure on information [74].

Data processing layer YARN and MapReduce establish two alternatives to complete information processing on Hadoop. They are intended to oversee work booking, assets, and the group. It merits seeing that YARN is more conventional than MapReduce.

(a) MapReduce

MapReduce is the scheduling framework, which is a system made out of a software framework and its execution. This function is the principal fundamental stride for the new age of Big Data control and examination apparatuses. MapReduce has a fascinating advantage for Big information solicitation. Generally, it works on handling enormous volumes of information through its productive and savvy instruments. It empowers to compose programs that can uphold parallel computing. Generally, the MapReduce scheduling framework utilizes 2 resulting capacities that handle information calculations: The Map work and the Reduce work. To be precise, a MapReduce program depends on the accompanying tasks:

- First, the Map work separates the information, such as long content records, into free information parcels that comprise key-esteem sets.
- Then, at that point, the MapReduce system sent all the key-esteem sets into the Mapper that measures every one of them independently, all over a few equal guide assignments across the bunch. Every information parcel is appointed to one of the processing nodes. The Mapper yields at least one middle-of-the-road key-esteem set. At this stage, the structure gathers all the transitional key-esteem sets and categorizes them by key. Therefore, the outcome is numerous keys with the list of related qualities.
- Then, the Reduce work is utilized to deal with the middle yield information. For every extraordinary key, the Reduce work combines the qualities related to the key as indicated by a predefined program (i.e., sifting, summing up, arranging, hashing, taking normal, or tracking down the most extreme). From that point onward, it produces at least one yield key-esteem set.
- At last, the MapReduce structure stores all the Key-esteem sets results in an output document. Inside MapReduce worldview, the NameNode executes as a JobTracker occasion to plan the various positions and convey undertakings over the slave node. The JobTracker screens the situation with the slave hubs and re-appoints assignments when they fizzle to protect implementation dependability. Every one of the slave nodes runs job tracker representatives for the allotted assignment. A TaskTracker occasion executes the assignments as indicated by the JobTracker and screens their execution. Every TaskTracker will utilize numerous Java Virtual Machines (JVMs) to accomplish a few guides or decrease the failures in the network. Typically, a Hadoop collection is made out of client providers, different DataNodes, and two NameNodes (essential and optional). The job of the client providers is first to stack information and afterward to submit MapReduce occupations to the NameNode. The essential Master NameNode is committed to facilitating and overseeing capacity and calculations. Then again, the auxiliary expert NameNode handles information replication and accessibility. An actual novel worker might deal with three jobs: customer, expert, and slaves in a little cluster within 40 nodes. Nonetheless, every job should be assigned to a solitary worker machine in medium and enormous groups.

(b) YARN

The YARN framework is more conventional than MapReduce. It gives superior versatility, equality, and progressed asset administration in contrast with MapReduce. It renders working framework capacities for Big Data scientific solicitation. The YARN Resource executive is integrated into the modified Hadoop structure. Specifically, the implementation of YARN takes place at the peak of HDFS. This position enables the parallel execution of multiple applications. It allows handling both batch processing and real-time interactive processing. YARN is compatible with the Application Programming Interface (API) of MapReduce. Users have just to recompile MapReduce jobs to run them on YARN. Unlike MapReduce, YARN enhances efficiency by splitting the two main functionalities of the JobTracker into two separate domains:

- The Resource Manager (RM) allocates and manages resources across the cluster.
- Application Master (AM) is a system comprised of huge archives that allocates specific tasks, which are equalized with TaskTrackers to track their enhancement. The AM empowers tasks such as accumulating books, perpetuating counters, self-detrainment failure, or latency.

(c) Cascading

A MapReduce structure is a rich Java API that gives a large number to quick and cost-efficient Big Data application advancement, testing, and reconciliation. Cascading possesses intriguing benefits, permits controlling the progressed questions, and takes care of complex work processes on Hadoop groups. It upholds adaptability, compactness, incorporation, and test-driven turn of events. This API adds a deliberation level on the highest point of Hadoop to improve on complex inquiries through a falling idea. The stacked information is handled and parted by a progression of capacities to get numerous flows called streams. Those streams structure non-cyclic coordinated charts and can be consolidated depending on the situation. The line gathering characterizes the stream to run between the information sources (Source Taps) and the yield information (Sink Taps) that are associated with the line. A line gathering might contain at least one Tuples of a given size. A cascading stream is compiled in Java and changed into exemplary MapReduce during execution. The streams are executed on Hadoop groups and depend on the accompanying cycle. A Flow example is a work process that first peruses the information from one or more Source Taps before measuring them by executing an assortment of equal or successive activities as characterized by the line group. Then, at that point, it composes the yield information into one or a few Sink Taps. A Tuple addresses a bunch of qualities (like a data set record of SQL table) that can be ordered with Fields and can be put away straightforwardly into any Hadoop File design as key/esteem pair. A tuple ought to have equivalent sorts to work with Tuple correlation. A huge number were added to the Cascading structure to upgrade its capacities, including:

- The pattern is used to fabricate prescient enormous information applications. It gives many AI calculations and empowers deciphering Predictive Model Markup Language (PMML) archives into solicitations on Hadoop.

- Scalding is utilized as a unique programming language to take care of practical issues. It depends on Scala language with straightforward arrangements. This expansion is constructed and kept up with by Twitter
- Cascalog permits boosting the application by utilizing Clojure (a powerful programming language dependent on Lisp lingo) or java. It upholds Ad-hoc inquiries by executing a progression of numerous MapReduce responsibilities to examine various sources (Local data, HDFS, and information bases). It gives a more significant level of reflection than Hive or Pig.
- Lingual gives an ANSI-SQL communication to Apache Hadoop and supports a quick relocation of information and jobs to and from Hadoop. Through Lingual, it is simpler to coordinate the current Business Intelligence instruments and different applications [47].

Information querying layer The advanced scripting language, Pig Latin, was developed by the non-proprietary system known as Hive, Pig, and JAQL. It diminishes MapReduce intricacy by supporting the parallel computation of MapReduce tasks and work processes on Hadoop. Pig-like Hive streamlines investigating and computing in equal monstrous informational indexes utilizing HDFS (e.g., complex information stream for ETL, different information examination) through its intuitive climate. Pig permits additional association with outer projects, like shell contents, pairs, and other programming dialects. Pig has its information framework known as Map Data, which is a bunch of key-esteem sets. Pig Latin has numerous advantages. It depends on a natural language structure to help simple MapReduce tasks and work processes (basic or settled streams). It lessens the advancement time while supporting similarities. Along these lines, clients can depend on Pig Latin language and a few administrators to transfer and handle information. Generally, administrators that cycle information establishes nodes in such DAC while edges introduce information streams. Despite SQL, Pig doesn't need composition and can handle semi-organized and unstructured information. It upholds a bigger number of information designs than Hive. A pig can run on both the nearby climate in a solitary JVM and the conveyed climate on a Hadoop group. JAQL is a decisive language on top of Hadoop that gives a question language and supports BDA [14, 24, 28, 45, 46]. It changes over significant level inquiries into MapReduce tasks. It was designed to inquire about semi-organized information dependent on JSON (JavaScript Object Notation) design. It is most often used to question different information organized just as numerous information types (e.g., XML, comma-isolated qualities (CSV) information, level documents). Along these lines, JAQL, like Pig, does not need information mapping. JAQL gives a few in-fabricated capacities, center administrators, and I/O connectors. Such components guarantee information preparing, stockpiling, making an interpretation of, and information changing over into JSON design. Apache Hive is an information stockroom framework intended to improve the utilization of Apache Hadoop. Rather than MapReduce, which controls information in the records through HDFS.

Hive empowers to address information in an organized dataset that mostly intimates to the clients. Generally, Hive's information model is founded on tables. Such tables address HDFS catalogs and are isolated into parts. Each segment is then partitioned

into pails known as buckets. Besides, Hive gives a SQL-like language called HiveQL that empowers clients to get to and control Hadoop-based information stockpiled in HDFS or HBase. In this manner, Hive is appropriate for some business applications. Hive is not suited for continuous exchanges. Indeed, it depends on low-dormancy activities. Like Hadoop, Hive is intended for huge scope computing, so even little positions might require minutes. In reality, HiveQL straightforwardly changes over inquiries (e.g., impromptu questions, joins, and rundown) into MapReduce occupations that are prepared as cluster failures. Hive also enables connecting conventional mappers and reducers when it is impossible or wasteful to communicate them in HiveQL. Contrary to SQL, which has a blueprint on-compose highlight, Hive has an on-read pattern and supports different compositions, which concedes the use of an outline until you attempt to peruse the information. However, the advantage is regarding faster stacks, and the disadvantage is that the questions are moderately slower. The hive needs full SQL support and does not give column-level embedded refreshes or erases. Thus, HBase is a better investment in this case [74].

Data access layer This section describes the data access layer, the data ingestion process, and the storage management process.

(a) Data ingestion: Sqoop, Flume, and Chukwa

1. Apache Sqoop in data ingestion

Apache scoop enables a command-line interface (CLI) that guarantees a proficient exchange of mass information between Apache Hadoop and organized data stores (like social data sets, endeavor information distribution centers, and NoSQL data sets). Sqoop offers many benefits, like quick execution, adaptation to internal failure, and ideal framework usage to lessen computation overhead to outer frameworks. The imported information is changed using MapReduce or some other significant level language such as Hive, Pig Pig, and JAQL. It permits simple reconciliation with HBase, Hive, and Oozie. At the point when Sqoop imports information from HDFS, the yield is provided in different documents. These records might be delimited content documents, parallel Avro, or SequenceFiles containing serialized information. The interaction of Sqoop Export peruses a bunch of delimited content documents from HDFS in equal, parse them into records, and supplement them as new lines in an objective data settable.

(2) The flume in data ingestion

The flume has a basic adaptable design and handles the flow of information streams. Flume depends on a basic extensible information model to deal with huge appropriated information sources and gives different components including adaptation to internal failure, tunable dependability system just as disappointment recuperation administration. However, the Flume supplements Hadoop well; a free segment can chip away at different stages. It is needed for its ability to run different cycles on a solitary machine. By utilizing Flume, clients can stream information from different and high-volume sources for ongoing investigation. Moreover, Flume gives an inquiry preparing motor that can change each new information bunch before it is directed to the predetermined sink.

(3) Chukwa in data ingestion

It is an information assortment framework based on the highest point of Hadoop. Chukwa probably screens enormous disseminated frameworks and utilizes HDFS to gather information from different information suppliers along with MapReduce to investigate the gathered information. It acquires Hadoop's adaptability and viability and establishes an interface to show screen, and dissect outcomes Chukwa empowers an adaptable and amazing stage for Big Data, which empowers examiners to gather and dissect big datasets just to screen and show results. Chukwa is organized as an assortment pipeline to guarantee adaptability, preparing stages just as characterized by the connection between stages. The framework depends on four fundamental segments: first, it depends on information specialists on each machine to transmit information. Authorities are then used to gather information from specialists and compose it into a steady stockpiling. MapReduce occupations are utilized to parse and file information. Clients can depend on an agreeable interface (HICC) to show results and information. It has an online interface style.

(4) Storm and spark storm

It is a non-proprietary dispersed framework that enjoys the benefit of taking care of continuous information handling conversely with Hadoop, intended for cluster computing. In contrast with the flume framework, storm demonstrates better productivity in carrying out perplexing computing prerequisites depending on Trident API. Strom depends on a network topology comprised of a total organization of streams, bolts, and spouts. A spout is a gathering of streams, and a bolt handles input flows to create output flows. Henceforth, the storm is appropriate to perform stream changes utilizing "spouts" and "bolts." The ISpout interface of the storm can uphold any approaching information. Indeed, by utilizing Storm, clients can ingest information from different continuous coordinated and offbeat frameworks. With the help of Bolts, Storm empowers to compose information to any yield framework. Storms gives the IBolt attachment that upholds any kind of yield framework, like JDBC (to store information to any social data set), Sequence Files, Hadoop segments, like HDFS, HBase, and other informing frameworks. The tempest bunch and Hadoop group are comparable.

Notwithstanding, it is feasible to run various geographies for various storm failures in Storm. Yet, in the Hadoop stage, the solitary alternative comprises carrying out Map-Reduce tasks for the relating applications. One fundamental distinction between MapReduce occupations and topologies is that when MapReduce stops working, a topology continues to compute the messages either constantly or until the client terminates. The storm is a simple to utilize, quick, adaptable, and flaw lenient framework; if at least one cycle comes up short, Storm will consequently restart it. The storm will reroute on the off chance that the cycle flops more than once. It may be utilized for some cases, such as ongoing examination, online AI, constant calculation, and appropriated RPC. The storm is utilized to plan results that would then be able to be broken down by other Hadoop devices. It can deal with million tuples each second. Storm gives an improved programming model like MapReduce, which shrouds the intricacy of creating conveyed applications. Flash resembles Hadoop, yet it depends on an in-memory framework to further develop execution. It is a perceived examination stage that guarantees quick,

simple, and adaptable processing. Flash handles complex examinations on huge informational collections. Surely, Spark runs programs up to 100 xs quicker than Hive and Apache Hadoop through MapReduce in-memory framework. Sparkle depends on the Apache Hive codebase. To further develop framework execution, Spark trade-out the actual execution motor of Hive.

Moreover, Spark offers APIs to help quickly advance applications in different dialects, including Java, Python, and Scala. Sparkle can work with all documents stockpiling frameworks that Hadoop upholds. Sparkle's information model depends on the Resilient Distributed Dataset (RDD) deliberation. RDDs establish a read-just assortment of items put away in framework memory across different machines. Such items are accessible without requiring plate access.

Moreover, they can be recreated if a parcel is lost. The Spark project comprises various parts for task planning, memory on the board, deficiency recuperation, and cooperating with capacity frameworks. The parts are recorded as follows:

5. Spark SQL

One significant component of Spark SQL is that it combines the two deliberations: social tables and RDD. Therefore, developers can ingest the data without much of a stretch and blend SQL orders to question outer informational collections with the complex investigation. Solidly, clients can run questions over both imported information from outer sources and information put away in subsisting RDDs. What's more, Spark SQL permits working RDDs or Parquet documents. It works with quick equal handling of information questions over enormous circulated informational collections for this reason. It utilizes inquiry dialects called HiveQL. For a quick application advancement, Spark has fostered the Catalyst system, which empowers clients through Spark SQL to add new advancements quickly.

6. Spark streaming

Spark flow is the supplemental part that gives programmed comparable, just as adaptable and liability un-biased streaming handling. It empowers clients to stream assignments by composing bunch-like cycles in Scala and Java. It is feasible to incorporate a bunch of occupations and intelligent questions. It implements each flow calculation as a progression of short bunch occupations on commemorative information put away in RDDs.

7. MLlib

MLlib is a conveyed AI system based on top of Spark. MLlib gives different upgraded AI calculations for execution, such as grouping, relapse, bunching, and cooperative separating. Like MLlib and Mahout, MLlib is valuable for AI classifications. They offer estimations for theme displaying and successive example mining. MLlib upholds additional relapse Models. Be that as it may, Mahout doesn't assist with such a framework. MLlib is generally youthful in contrast with Mahout.

8. GraphX

GraphX comprises a library for controlling charts and executing diagram equal calculations. Like Spark SQL, GraphX, and Spark flowing expands the elements of Spark. Consequently, clients can make a coordinated chart with self-assertive assets joined to every edge and vertex. GraphX offers various administrators that help chart control. It gives likewise a library of diagram calculations (triangle tallying).

Storage administration

HCatalog apache

HCatalog imparts the table and the repository board administration for Hadoop clients, which empowers interoperability among the information handling apparatuses. HCatalog accomplishes this through a common pattern and information type instruments, which provide an interface to improve peruse and compose information activities for any information design for which a Hive SerDe (serializer-deserializer) can be composed. The framework administer imparts the SerDe, OutputFormat, and InputFormat. The disconnected table of HCatalog gives a social perspective on information in HDFS and permits us to see dissimilar information designs in an even arrangement. Therefore, the clients are not required to know the location or method of information storage. Besides, HCatalog upholds clients with different administrations and informs information accessibility, which gives a REST interface to allow admittance to Hive Data Definition Language (DDL) tasks. It additionally gives warning services that tell work process devices (like Oozie) when new information opens up in the stockroom [74].

Review and analysis of big data technologies

This section reviews and analyzes 37 articles related to big data technologies. Using these articles, this section reviews the application of big data in various fields. Big data is applied in almost all fields, such as smart cities, network communication, business management, smart home, transport, sentiment analysis, decision support or opinion mining, smart home, privacy, health care, industrial application, and agriculture. The big data application uses various techniques such as machine learning, deep learning, Cloud computing techniques, Edge computing techniques, concentric computing, and IoT. A brief description of is application of Big data is illustrated below. Figure 1 depicts the diagrammatic representation of the big data application. This section briefly analyzes some of the existing articles related to big data applications, with their methods, advantages, and disadvantages.

Review based on big data applications

This section illustrates the review of the application of big data in various fields. The use of big data is prevalent in almost all industries, such as smart cities, network communication, business management, smart home, transport, sentiment analysis, decision support or opinion mining, smart home, privacy, health care, industrial application, and agriculture. A brief description of the application of big data is illustrated below. Figure 1 depicts the diagrammatic representation of the big data application.

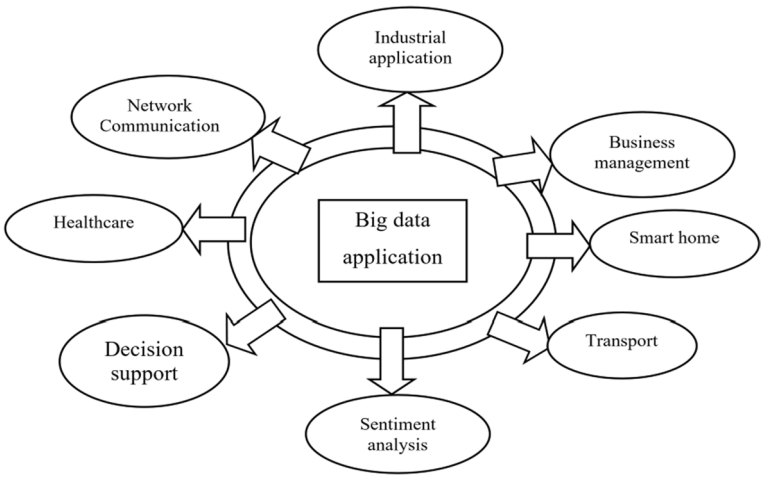

Fig.1 Big data application

Machine learning techniques

Kibria et al. [52] suggested information-driven cutting-edge remote organizations, where the organization administrators utilize progressed information investigation, AI, and machine learning approaches. This method evaluates the information sources and solid drivers for the establishment of the information examination and the job of ML, man-made consciousness in making the framework smart regarding acting naturally mindful, self-versatile, proactive, and prescriptive. Concerning information examination, several organizational and streamlining plans have been introduced.

Raginia et al. [58] presented an enormous information-driven methodology and described the different stages associated with estimation order. The principle commitment is the point-by-point investigation of the learning and arrangement strategy to order the requirements of individuals during the hours of the catastrophe. To examine the assumptions about the different necessities of individuals influenced by the fiasco, and suggests a technique to imagine the opinion. The blend of abstract expression and AI algorithm provides better grouping exactness for calamity information. The large information-based examination likewise dissects the different difficulties associated with utilizing Twitter for calamity reaction and recuperation. It is clear from the close investigation that the vocabulary accessible for text examination should be expanded to accommodate catastrophe-related information. Likewise, building metaphysics for emergency information can work on grouping individuals' necessities during calamity.

Deep learning techniques

Raut et al. [53] presented hybrid Structural Equation Modeling based on the ANN model utilized to evaluate 316 reactions of Indian expert specialists. The factor examination results demonstrate that administration and initiative style, state and government strategy, provider integration, interior business cycle, and integration of client affect huge information investigation and supportability rehearses. Besides, the outcomes from the underlying conditions demonstrated were incorporated into the Artificial Neural Network organization model. The investigation discoveries show that administration and authority style state and local government strategy as the two most significant indicators

of enormous information examination and manageability rehearses. The outcomes give one-of-a-kind bits of knowledge into assembling firms to further develop their supportable business execution from a task executive's perspective.

Cloud computing

Lu and Xu [54] presented a conventional framework design for cloud-dependent assembling hardware for digital production frameworks. A huge information examination is presented, permitting fabricating gear to be associated with the cloud and made accessible for arranging on-request manufacturing administrations. An industry execution in a world-driving apparatus arrangement supplier affirms that the presented framework design for cloud-based assembling gear can effectively empower on-request fabricating administrations provisioned using the Internet and can be expanded to organizations that are undertaking to change inheritance creation frameworks into cloud-based digital frameworks.

Yassine et al. [61] presented a stage for intelligent home IoT enormous information investigation with loud and distributed computing. This method provides prerequisite evaluation and representation of the stage segments and establishes the path toward playing out the evaluation in the fog node. The outcomes show the potential uses of the framework from various angles. For instance, utilizations of the information procured may incorporate movement acknowledgment to recognize medical conditions, distinguish energy utilization examples and energy-saving arranging, and guarantee proficient tasks according to the perspective of energy utilization.

Shorfuzzaman et al. [62] developed a smart cloud-based processing system that can successfully examine versatile students' enormous information and offer this information continuously to work with compelling dynamics in advanced educational organizations. The cloud-depended structure beats the restricted handling capacity of cell phones by offloading certain hefty computational parts to the cloud, which can give sufficient calculation and capacity assets. Specifically, this method focuses on the particular use of huge information investigation procedures in a versatile learning structure to work with its plan choices. In conclusion, this method addresses the issue of students' preparation for versatile learning reception in advanced education establishments by investigating the rousing variables to embrace this arising innovation utilizing a drawn-out conjectured innovation acknowledgment model.

Rajeswari et al. [73] suggested an advanced model for the agricultural sector based on cloud-computing techniques is to foresee the harvest yield and choose the better harvest succession dependent on the past crop arrangement in similar farmland with the dirt supplement current data. Through continuous examining of soil, the rancher will want to get current compost necessities for the farming harvest. This is a fundamental prerequisite for the farming area in India to get further developed yield creation with a decrease in the expense of manure necessities keeping soil with wellbeing flawless. As the information is gathered throughout the years for crop subtleties and soil conditions, this model investigates the big data to find the best edit arrangement, to determine the next harvest to be developed for better creation, to complete yield creation in the space of revenue, to determine the all-out compost necessities, and ascertain other information of the premium.

Fog computing

Darwish et al. [63] presented the design for an on-time ITS enormous information investigation in the IoV environment. The design combines three measurements: smart processing (for example, cloud and mist processing), continuous enormous information investigation, and IoV measurement. In addition, this research gives an exhaustive portrayal of the nature of IoV, the ITS enormous information qualities, the lambda engineering for continuous huge information examination, and a few clever figuring advancements. All the more significantly, this research examines the chances and difficulties that face the execution of mist processing and continuous enormous information investigation in the IoV.

Edge computing

Garg et al. [66] presented a high-level vehicular correspondence method where RSUs are restored by edge registering stages. Then, at that point, secure V2V and V2E correspondence are planned to utilize the Quotient channel, a probabilistic information structure. In synopsis, an intelligent security system for Vehicular Ad hock Networks (VANETs) outfitted with edge registering nodes and 5G innovation was designed to upgrade the abilities of correspondence and calculation in the cutting edge smart city environments. The experimental outcomes exhibit that the model beats the ordinary vehicular models by furnishing an energy-proficient secure framework with the least delay.

Concentric computing

The survey on the smart city-application-dependent big data application is enumerated below. Shahat Osman [51] presented three-layer frameworks known as Smart City Data Analytics Panel (SCDAP) for the smart city application. The framework consists of three layers: data computing, security, and platform. In establishing a layered framework approach, the accession of the normalized information, empowering of both ongoing and verifiable information investigation, the backing of both iterative and consecutive information preparation, and versatility are the usually followed plan standards. Furthermore, on a level plane, adaptable stages, cloud frameworks, information mining, AI, in-memory datasets, and information perception are the empowering innovation. On the opposite side, two extra capacities to the examination structures for brilliant urban communities, to be specific: model administration and model accumulation, are the fundamental commitment of the model. SCDAP presents new functionalities to large information examination systems for smart city applications. The primary component of this engineering is restricted to the Apache Hadoop suite as fundamental information stockpiling and the executives' layer. The partition between SCDAP functionalities and the basic information stockpiling and the board layer improves the consensus of SCDAP and its capacity to manage numerous different stages.

Rehman et al. [68] presented an effective distribution system of the Concentric Computing Model (CCM) for the application BDA to work on functional productivity and application execution in IIoT-depended frameworks. This model featured the exhibition and correspondence destinations that can be met through taking on CCM in IIoT.

IoT

Al-Ali et al. [55] presented an Energy Management System (EMS) for home smart home applications. In this framework, each home gadget is interfaced with an information securing module that is an IoT object with an interesting IP address bringing about a huge cross-section remote organization of the device. The information obtaining System on Chip (SoC) module gathers energy utilization information from every savvy home gadget. It communicates the information to a unified worker for additional handling and examination. This data from all local locations collects by the utility's distributors as Big Data [55].

Rathore et al. [59] suggested a framework for a Smart Digitalized City that utilizes IoT devices to gather information about the city and Big Data investigation for information procurement. Advanced Digital City engineering and execution model is used to carry out the framework that can deal with an enormous measure of city information and offer direction to the metropolitan specialists to make their regions more intelligent and computerized. The framework execution involves different advances, including information age and assortment, amassing, filtration, arrangement, preprocessing, processing, and dynamic. The gathered information from all smart frameworks continuously handled to accomplish smart city utilizing Hadoop working under Apache Spark. Existing datasets produced by smart homes, savvy parking, climate frameworks, contamination checking sensors, and vehicle network are utilized for investigation and testing. All the datasets are replayed alongside the genuine execution of the vehicular organization to take the constant information to test the framework. This method attempted to take care of the greater part of the issues that a normal resident appearances and likewise gave access to the public authority to take savvy choices at constant. Furthermore, the model helps to restrain queries and issues generated by residents. STB utilized a diagram approach executed using huge chart handling devices. The outcomes demonstrate that the framework is more productive, versatile, and fit for working in an ongoing climate.

Big data analytics

Amini et al. [57] presented a complete and adaptable design for ongoing traffic lights dependent on Big Data investigation. The framework depends on a methodical examination of the necessities of the region. The framework has been reified in the model a stage utilizing Kafka and has been utilized in working a criticism control circle to open or close the hard shoulder of a road. The fundamental restriction of the investigation was the absence of admittance to genuine information.

Yacchirema et al. [64] presented a framework dependent on 3-level architecture to assist continuous observation of Obstructive sleep apnea (OSA) in older individuals, and directing their treatment has been established and carried out. The framework is executed utilizing the heterogeneous and non-meddlesome device, IoT conventions, segments of standard stages, low-power innovations, huge information advances, and haze and Cloud registering approaches. Exploiting the Big Data instruments cluster information handling at the Cloud layer has been carried out. It can play out a graphic investigation that measurably subtleties the conduct of the information from the intelligent IoT door and foresee the most un-structured region dependent on the toxin's information accessible in smart urban communities to direct the diagnosis of

OSA. The examined information is conveyed to a worker, which shows the data in a Web Indian University IU, so the medical services experts engaged with the consideration of the old individuals can without much of stretch access from anyplace whenever and from any gadget.

Gohar et al. [65] exhibited a Big Data evaluation strategy for ITS, which dissects and stores the ITS information proficiently and assists the city organization in settling on better choices dependent on the sort of representations accessible. The strategy has an underlying stockpiling and investigation ability to work with ITS information and is made out of four modules: (1) Big Data Acquisition and Preprocessing Unit, (2) Big Data Processing Unit, (3) Big Data Analytics Unit, and (4) Data Visualization Unit. This method is assessed utilizing Hadoop, subsequently approving the confirmation of the idea.

Wang et al. [71] presented a framework level model for dependability assessment in the Energy Internet on a basic energy capacity to investigate the little aggravation solidness region (SDSR), where SDSR can be obtained by assessing the functional information edge of appropriated ages (DGs). The edge is assessed based on energy utilization instead of balance nodes, which utilize the energy work hypothesis and lessens the calculation intricacy. In addition, the enormous information added in the model approximates evaluation calculation into the hyper-plane fitting to advance and break down the SDSR.

Miah et al. [58] presented a strategy to remove, rank, recognize and distinguish significant traveler data from unstructured huge informational indexes for supporting the Database Migration Option (DMO) dynamic decision making. By investigating geotagged photographs with other related subtleties, our strategy is material to various objections and produced valuable outcomes, as outlined for the instance of Melbourne, Australia. This method followed the setup hypothesis and methodological rules of Data Subject Request (DSR) for the plan, advancement, and spread of the generated noise.

Hadoop technologies

Babar et al. [69] presented a Hadoop-dependent design to manage big data stacking and preparing, and the administration design consists of two distinct modules. The exhibition and proficiency of information stacking are tested to develop a tweaked methodology for stacking Big Data to a conveyed and preparing stage. Information stacking is performed and contrasted over and over against various choices to analyze information ingestion into Hadoop. The experimental outcomes are recorded for different traits alongside manual and customary information stacking to feature the organization's proficiency. Then again, the handling is accomplished utilizing YARN bunch the board system with explicit customization of dynamic planning. Additionally, the method's suitability in terms of planning and computation is highlighted and embellished in terms of throughput.

Other techniques

Chang [72] suggested a model for the Social Network Analysis Platform (SNAP), which demonstrates how to extricate the information from Facebook within the established Social Network API, which evaluates Big Data of the correspondence network. The description of the organization, as well as the methods to dissect and imagine information, have been elaborated. The six elements of Social Network API can be utilized to dissect the information that explains the connection between various organizations. They can show that SNAP can be utilized as a Customer relationship management (CRM), Enterprise resource planning (ERP), and probabilistic management information system (MIS) stage to separate data of the organizations and perform a profound investigation of each conceivable circumstance. Furthermore, six elements of Social Network API have been tried widely with huge scope reenactments. The results support the versatility of our SNAP proposition, with the goal that the administration can be up and running when there are 50,000 simultaneous clients or solicitations, which can further test the speed and exactness of big data informal community investigation.

Analysis and discussion

This section illustrates the evaluation and discussion of methods, metrics, and tools used in the research papers. The analysis results indicate the importance of the number of implementation strategies based on big data.

Analysis based on the utilized techniques

Table 1 illustrates the analysis of research papers in terms of the techniques utilized in the big data application. The review papers commonly utilized the database, such as machine learning, Deep-learning, cloud computing, and big data analytics methods. Big data analytics is the commonly utilized method among the different applications. Among the 25 papers, 9 papers utilize the BDA, 4 papers utilize machine learning techniques, 4 papers utilize cloud computing, and 1 paper uses the Hadoop framework. Figure 2 depicts the chart based on the analysis of the techniques.

Table 1 Analysis based on utilized methods

Techniques	Papers
Machine learning	[29, 52, 60, 67]
Deep learning	[53, 67]
Cloud computing	[54, 61, 62, 73]
Fog	[61, 63]
IoT	[55, 59]
Big data analytics	[29, 55–60, 65, 71]
Edge computing	[66]
Concentric computing	[51, 68]
Hadoop	[69]
Other techniques	[51, 70, 72]

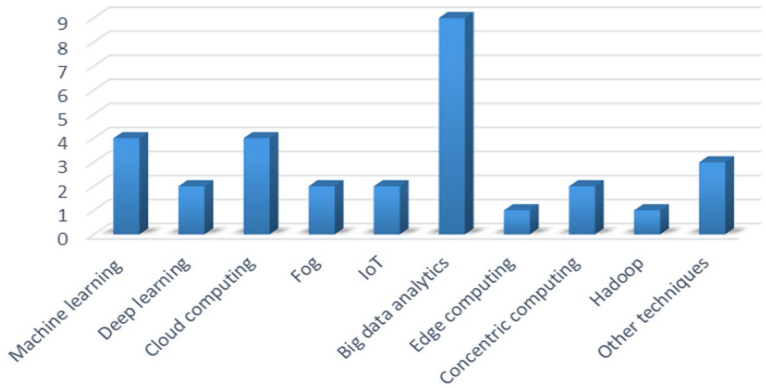

Fig.2 Chart based on the techniques used in the big data applications

Table 2 Analysis based on the performance metrics

Performance metrics	Papers
Accuracy	[51, 60, 64]
Resource utilization	[52]
RMSE	[54]
SD	[54, 62, 67]
Scalability	[55]
Speed	[55]
Security	[55]
Processing time	[29, 60, 69, 72]
TPR	[58]
FPR	[58, 60, 66]
Throughput	[58, 69]
Detection rate	[60, 66],
Energy consumption	[61]
Mean	[62]
Item loading	[62]
Delay	[66]
Stability	[71]
SINR	[67]

Analysis based on the performance metrics

Table 2 demonstrates the evaluation of the research articles in terms of performance indices utilized in big data applications. The accuracy, resource utilization, RMSE, speed, scalability speed, processing time, Peak Signal-to-Noise Ratio (PSNR), False Acceptance Ratio (FAR), and False Rejection Rate (FRR), delay, and stability are the commonly utilized performance metrics. Accuracy is the widely-accepted parameter utilized for evaluating the big data application. Figure 3 illustrates a chart based on the analysis of the performance metrics.

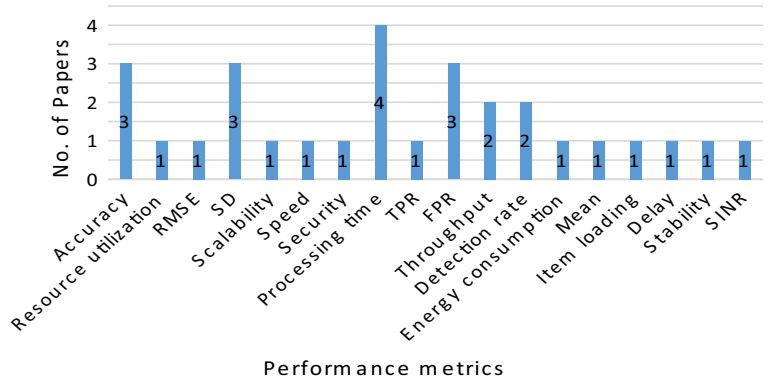

Fig. 3 Chart based on the analysis of the performance metrics

Table 3 Analysis based on year of publication

Year	Paper
2015	NIL
2016	[58]
2017	[55, 57, 59, 70–73]
2018	[51–54, 58, 60–65, 68]
2019	[66, 67, 69]
2020	[29, 56]

Table 4 Analysis based on journals

Journal	Research articles
Elsevier	[51, 53, 54, 56, 58, 59, 61, 62, 65, 69, 70, 72]
IEEE	[60, 73]
Springer	[29, 52, 55, 57, 63, 64, 66–68] and [71]
Wiley online library	[75–77, 85] and [74]
International journal of physics	[78–84] and [85]

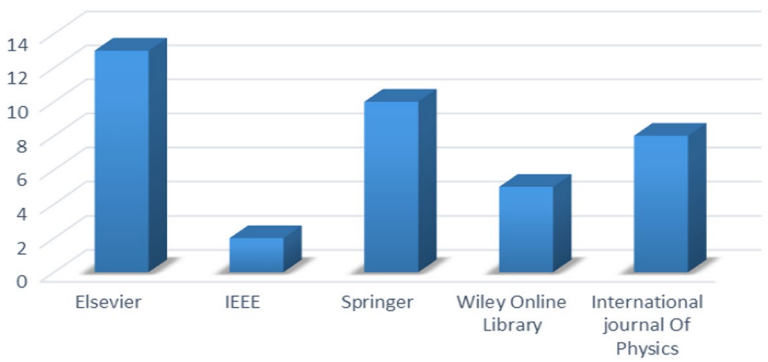

Fig. 4 Chart based on the journal publications

Analysis based on the publication years and journals

Table 3 contains a summary of the research papers' evaluations based on the year of publication. Most of the papers were published in 2018, 2019, and 2020. The journal publication details are enumerated in Table 4 and Fig. 4, respectively. Most of the articles are published in standard Journals and below the current years.

Challenges associated with big data applications

- Most big data analytic system is limited to the Apache Hadoop suite. Hence, the separation of the data storage layer and the management layer is a challenge. The separation between the mentioned layers will increase the functionality of the big data application.
- Balancing the distributed and the centralized functionality is the main issue experienced in the big data application based on deep learning [49].

Future direction

- Requires ensemble algorithm for effectively classifying the storage and the management layer.
- Export or the import function needs to be enabled to increase the functionality of the system
- We need to develop a dynamic edge computing method to establish a balance between centralized and distributed functionality.

Big data analytics: issues and challenges

- Big Data is confronting various difficulties in managing the organization [12, 15, 16, 33]. The framework working with big data needs to comprehend the requirement of innovation and the client's requirement. Meeting the difficulties introduced by big data is troublesome, and the volume of information expanding each day, speed of its age is expanding quicker than at any time in recent memory; an assortment of information is likewise growing. Current instruments, advances, engineering, executives, and investigation approaches can't copup with the intricacy of information introduced. A few difficulties are described below.
- Protection, Security, and Trust: organization utilizing big data resolved to ensure the protection and security of its clients and ought to guarantee that the association should agree on all protection and security-related demonstrations to improve the put down clear stopping points for the use of individual data [22]. The confidence in the association should increase with the volume of incrementing information. The trust clients have in these organizations and their capacities to safely hold an individual's data can undoubtedly be influenced by the spillage of information or data into public space.
- Information Management and Sharing: the agencies understand that for information to have any worth, it should be discoverable, open, and usable. Organizations should

accomplish these necessities yet hold fast to security laws. The latest things towards open information have emphasized making datasets accessible to general society. Offices should put center around making information accessible, open, and normalized inside and between organizations so that it permits offices to utilize and team up to the degree made conceivable by protection laws.

- Innovation and Analytical abilities: big data analytics put a parcel of weight on Information and Communication Technology (ICT) suppliers for growing new apparatuses and innovation to deal with complex information. The current instruments and innovations can't store, measure, and break-down huge measures of assorted information. The sellers and engineers of big data frameworks and arrangements, including open source programming, are becoming more competent to improve on the difficulties of big data analytics. Some specific challenges [48] related to Big Data and Analytics are:
- Information Storage and Retrieval: currently, accessible advancements can deal with information sections and information stockpiling. Yet, the apparatuses intended for the exchange process can search for a small amount of data rather than an enormous pile of information. The most effective method to deal with semi or unstructured information for handling is yet obscured [2].
- Information Growth and Expansion: as the associations build their administrations, their information is expected to develop. The association considers information development due to information filled in lavishness and information advanced with new strategies [2].
- Speed and scale: it is hard to acquire knowledge into information within the period when there is an enhancement in the volume of data. Acquiring knowledge about information is a higher priority than handling the total arrangement of information. Computing close to continuous information will consistently require preparing stretch to create acceptable outcomes [2].
- Organized and unstructured information: the transition between organized information stockpiled in the clear-cut tables and unstructured information (pictures, recordings, text) needed for examination will influence the preparation of information from beginning to end. Information portrayal and computing will become more adaptable with the development of new non-relative innovations [2].
- Information proprietorship: very tremendous measure of information dwells in the workers of web-based media specialist organizations. They do not possess this information; however, they store information about their clients. The genuine proprietor of the page has made the page or record. Information proprietorship is a growing challenge in the online media space [2, 18].
- Data Skew: in a dataset, the uneven distribution of data is named data skew, which is valid only in Parallel Processing architecture where Data Distributed Processing occurs. The importance of data distributed processing is that instead of a single processor doing a job, if this job is divided into multiple parallel smaller jobs processed by the different processors, that job as a whole is completed faster. This reduces the time of execution and improves performance [90].
- Data Variety: one of the key components of Big Data is data variety. Combining data from several sources and dispersing the data in different ways leads to data variety.

The use of processing resources, such as CPU use, varies greatly due to this aspect of big data. DVFS is a technique utilized to reduce the issues and consume the energy of the CPU [91].

- Limited resources: edge computing is a distributed computing paradigm that brings computation and data storage closer to the data sources. It is expected to improve response time and save bandwidth [92].
- Edge Data Processing: approximation approaches in Big Data can be useful for applications that examine incoming data, logs, and queries to produce aggregated results or dashboards. These applications use the most energy, money, and time possible because the output is significantly smaller than the input [93].
- Consumption of energy: data centers vary in efficiency, with a significant proportion of power delivered to a data center may be used for cooling and other ancillary functions rather than delivered to the IT equipment. Utilization rates should be high in data centers where work can be scheduled. Maintaining high utilization in other areas is more problematic unless demand is very predictable [94].

Comparative with existing survey method

Oussous et al. [86] review the recent technologies employed in big data analytics, which helps to select the appropriate technologies to meet the technical requirements and the application. The survey provides a comparison between the different storage layers, such as the management layer, data querying layer, data processing layer, management layer, and the data access layer. It provides brief descriptions of big data's advantages, limitations, and technological features. This survey article explains the significance and taxonomy of big data. The article also explains the tools utilized for big data processing. Further, the paper includes the analysis based on the utilized methods, year of publication, and performance metrics, which adds value to our survey article compared to the existing studies.

Chen et al. [87] provide a brief overview of big data issues with current technologies, techniques, and opportunities. The survey article provides descriptions of quantum computing, cloud computing, and biological computing. However, this article fails to analyze the current tools utilized in big data technologies, such as storm, spark, and Hive. Our survey article analyzes the latest tools with their merits and demerits. Further, our paper provides a brief explanation of the Hadoop ecosystem.

By considering various existing methods, the analysis based on big data under various criteria like techniques, tools, application, achievements, advantages, and disadvantages is presented in Table 5.

Conclusion

This article discusses big data terminologies, applications, and technologies and enumerates the issues associated with big data applications. This survey explains the significance of big data, its taxonomy, and details the basic terminologies used in big data. Moreover, the survey describes the technologies used in big data applications and the various complexities and challenges of big data. The paper focuses on the various techniques presented in the literature to restrain the issues related to big data. The big data

Table 5 Comparative analysis based on various criteria

Ref	Technique	Advantage	Disadvantage	Achievements	Tools	Application
[95]	Selection of optimal features for the applications of big data	It is efficient in the reduction of obstacles present in the feature selection model	The rate of convergence is low	Compared to other utilized algorithms' the accuracy levels of PSO and GWO are 86.8 and 81.6, hence obtained the average accuracy is 90 percent	MATLAB	Data mining
[46]	Analysis based on middleware to overcome the performance issues using machine learning	In controlled environments, establish realistic application workloads	The index will use more memory and slow write operations to maintain the secondary index			Data-intensive application
[96]	To provide better quantity service for intensive data applications in a cloud environment using Aneka	Minimum usage of resources while computing the risk task	Cost of scheduling is high		Basic Local. Alignment Search Tool (BLAST)	data-intensive application
[75]	Coreset-based data prioritizing solution to address security challenges brought on by jamming attacks	Improved quality of cluster with high detection efficiency	Inability to recover from database corruption	Based on the analysis of the quality index, the method obtained 0.805 using the Davies-Bouldin index (DBI)	NS-2	VANET
[98]	The use of the Non-dominated Sorting Differential Evolution (NSDE) algorithm—to increase the general superiority of the placement methods	The method achieves better load balancing and power consumption	Reduced latency for storage systems		MacBook Pro 2019	IoT-based power electronic applications
[99]	Accelerating computing and data mining operations in the cloud	Provides highly efficient approximate computing	Fails to contribute to reaching the highest speed and complexity	Hadoop speedup of 17 × and Sappox speedup of 8 × with 5% error tolerance	CUDA	Applications of Data mining
[23]	The performance of data-intensive applications using unstructured data using the Spark framework for hybrid program analysis	Improved performance on optimized code	Low performance due to the maximum size of data	The application can be accelerated by 7.47% and 2.967% more quickly with EP and CM		Data-intensive application

Table 5 (continued)

Ref	Technique	Advantage	Disadvantage	Achievements	Tools	Application
[90]	Workload allocation for data-intensive applications using the Energy Efficient and Bandwidth-Aware (EEBA)	Improved QoS and workload	High computation time	High cloud QoS with 11%, 38%, and 15% rate of makespan augmentation using Simple, Mixed, and Heavy BoT	CloudSim	application with a high data volume
[100]	Azure cloud analysis of physics data intensively	Increased efficiency and quality of research	High collection of data utilized with high computing time	For CPU doubling the CMS UCSD resources deliver a 7 M core, and the cloud-enabled CE provides 7.3 M hours each month	Grid Community Toolkit (GCT)	Data-intensive application
[30]	Spark Streaming Analysis for Data-Intensive Pipelines	Provides high throughput with quick efficiency of data transfer	High memory consumption	SUM Statistics for the GC The throughput of the Server Application with Backpressure is 100%	GC Analyser	Data-intensive application
[37]	A cross-point array dubbed with XAM serves as the foundation of Monarch, a resistive 3D stacked memory	Cost efficiency with improved performance	High latency	In String-Match, the better performance of Monarch occurs 24 × , 11 × , 12 × , and 14 × in HBM-SP, CMOS, HBM-C, and RRAM	ESESC multicore simulator	memory-intensive applications
[88]	To bridge the gap between programming models, HPC languages, and Big Data using the IgnisHPC5 framework	Improved performance with efficient execution	Lack of resources	MSAProbs produce a maximum performance difference of 0.4 percent	Python	MPI based applications

techniques are analyzed under several factors, including the year of publication, performance metrics, achievement of the existing model, and methods utilized.

Furthermore, this research article presents an evaluation of the methods in terms of their advantages and disadvantages. This review paper then discusses the big data opportunities and applications and provides a detailed sketch of the big data processing frameworks, like Hadoop, MapReduce, and Spark. Finally, the paper describes the future direction of research. Thus, the review summarizes the background of big data, its terminologies, techniques, and applications and focuses on the correct direction for further research. However, our research article fails to provide a deep insight into techniques such as deep learning, machine learning, and data visualization, which will be considered in future research.

Acknowledgements
The authors gratefully acknowledge the financial support from Wenzhou-Kean University.

Author contributions
Hemn contributed to the design and implementation of the research, the analysis of the results, and the manuscript's writing and contributed to the overall direction.

Funding
This work was supported by Leading Talents of Provincial Colleges and Universities, Zhejiang-China (#WB20200915000043).

Availability of data and materials
All data based on references.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
The authors give the Publisher permission to publish the work.

Competing interests
The authors declare that they have no competing interests.

Received: 25 February 2022 Accepted: 24 October 2022
Published online: 17 November 2022

References

1. Wei J, Chen M, Wang L, Ren P, Lei Y, Qu Y, Jiang Q, Dong X, Wu W, Wang Q, Zhang K. Status, challenges and trends of data-intensive supercomputing. CCF Transactions on High Performance Computing. 2022. 1–20.
2. Zakir J, Seymour T, Berg K. Big data analytics. Issues Inf Syst. 2015;16:2.
3. Margara A. A unifying model for distributed data-intensive systems. In Proceedings of the 16th ACM International Conference on Distributed and Event-Based Systems. 2022. p. 176–179.
4. Hu H, Wen Y, Chua TS, Li X. Toward scalable systems for big data analytics: a technology tutorial. IEEE Access. 2014;2:652–87.
5. Fisher D, DeLine R, Czerwinski M, Drucker S. Interactions with big data analytics. Interactions. 2012;19(3):50–9.
6. Silva BN, Diyan M, Han K. Big data analytics. In: Khan M, Jan B, Farman H, editors. Deep learning: convergence to big data analytics. Singapore: Springer Singapore; 2019. p. 13–30.
7. Vassakis K, Petrakis E, Kopanakis I. Big data analytics: applications, prospects and challenges. In: Skourletopoulos G, Mastorakis G, Mavromoustakis CX, Dobre C, Pallis E, editors. Mobile big data. Cham: Springer International Publishing; 2018. p. 3–20.
8. Kwon O, Lee N, Shin B. Data quality management, data usage experience and acquisition intention of big data analytics. Int J Inf Manage. 2014;34(3):387–94.
9. Elgendy N, Elragal A. Big data analytics in support of the decision making process. Procedia Comput Sci. 2016;100:1071–84.
10. Gandomi A, Haider M. Beyond the hype: big data concepts, methods, and analytics. Int J Inf Manage. 2015;35(2):137–44.
11. Vitale CH, Kennedy ML, Ruttenberg J. Advancing data science, data-intensive research, and its understanding through collaboration. In: Mani NS, Cawley MA, editors. Handbook of research on academic libraries as partners in data science ecosystems. Hershey: IGI Global; 2022. p. 25–44.

12. Zhou ZH, Chawla NV, Jin Y, Williams GJ. Big data opportunities and challenges: discussions from data analytics perspectives [discussion forum]. IEEE Comput Intell Mag. 2014;9(4):62–74.
13. Wamba SF, Gunasekaran A, Akter S, Ren SJF, Dubey R, Childe SJ. Big data analytics and firm performance: effects of dynamic capabilities. J Bus Res. 2017;70:356–65.
14. Mohamed A, Najafabadi MK, Wah YB, Zaman EAK, Maskat R. The state of the art and taxonomy of big data analytics: view from new big data framework. Artif Intell Rev. 2020;53(2):989–1037.
15. Lv Z, Song H, Basanta-Val P, Steed A, Jo M. Next-generation big data analytics: state of the art, challenges, and future research topics. IEEE Trans Industr Inf. 2017;13(4):1891–9.
16. Ardagna CA, Ceravolo P, Damiani E. Big data analytics as-a-service: Issues and challenges. In proceedings of IEEE international conference on big data (big data). 2016. p. 3638–3644.
17. Jun SW, Liu M, Lee S, Hicks J, Ankcorn J, King M, Xu S. Bluedbm an appliance for big data analytics. In: Marr D, editor. ACM/IEEE 42nd annual international symposium on computer architecture ISCA. Manhattan: IEEE; 2015. p. 1–13.
18. Chandarana P, Vijayalakshmi M. Big data analytics frameworks, In proceedings of IEEE International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA). 2014. p. 430–434.
19. Gupta R, Gupta H, Mohania M. Cloud computing and big data analytics: what is new from databases perspective? Springer: In proceedings of international conference on big data analytics; 2012. p. 42–61.
20. Salloum S, Dautov R, Chen X, Peng PX, Huang JZ. Big data analytics on apache spark. Int J Data Sci Anal. 2016;1(3):145–64.
21. Erdei RM, Toka L. Optimal Resource Provisioning for Data-intensive Microservices. In NOMS 2022–2022 IEEE/IFIP Network Operations and Management Symposium. IEEE. 2022. (pp. 1–6).
22. Jensen, M., "Challenges of privacy protection in big data analytics", In proceedings of IEEE international congress on big data. 2013. p. 235–238.
23. Wang WY, Yin J, Chai Z, Chen X, Zhao W, Lu J, Sun F, Jia Q, Gao X, Tang B, Hui X. Big data-assisted digital twins for the smart design and manufacturing of advanced materials: from atoms to products. J Mater Inform. 2022;2(1):1.
24. Iqbal R, Doctor F, More B, Mahmud S, Yousuf U. Big data analytics: computational intelligence techniques and application areas. Technol Forecast Soc Chang. 2020;153:119253.
25. Ardagna CA, Bellandi V, Bezzi M, Ceravolo P, Damiani E, Hebert C. Model-based big data analytics-as-a-service: take big data to the next level. IEEE Trans Serv Comput. 2018;14(2):516–29.
26. Shu H. Big data analytics: six techniques. Geo-spatial Inf Sci. 2016;19(2):119–28.
27. Cevher V, Becker S, Schmidt M. Convex optimization for big data: scalable, randomized, and parallel algorithms for big data analytics. IEEE Signal Process Mag. 2014;31(5):32–43.
28. Barba-González C, García-Nieto J, del Mar Roldán-García M, Navas-Delgado I, Nebro AJ, Aldana-Montes JF. BIGOWL: knowledge centered big data analytics. Expert Syst Appl. 2019;115:543–56.
29. El-Hasnony IM, Barakat SI, Elhoseny M, Mostafa RR. Improved feature selection model for big data analytics. IEEE Access. 2020;8:66989–7004.
30. Matteussi KJ, dos Anjos J, Leithardt VR, Geyer CF. Performance evaluation analysis of spark streaming backpressure for data-intensive pipelines. Sensors. 2022;22(13):4756.
31. Ortega F, Cano EL. Sensor data analytics: challenges and methods for data-intensive applications. Entropy. 2022;24(7):850.
32. Veiga J, Expósito RR, Pardo XC, Taboada GL, Tourifio J. Performance evaluation of big data frameworks for large-scale data analytics. In proceedings of IEEE International Conference on Big Data (Big Data). 2016. p. 424–431.
33. Venkatesh K, Ali MJS, Nithiyanandam N, Rajesh M. Challenges and research disputes and tools in big data analytics. Int J Eng Adv Technol. 2019;6:1949–52.
34. Divya KS, Bhargavi P, Jyothi S. Machine learning algorithms in big data analytics. Int J Comput Sci Eng. 2018;6(1):63–70.
35. Chawda RK, Thakur G. Big data and advanced analytics tools. In proceedings of symposium on colossal data analysis and networking (CDAN). 2016. p. 1–8
36. Wilcox T, Jin N, Flach P, Thumim J. A big data platform for smart meter data analytics. Comput Ind. 2019;105:250–9.
37. Prasad AK, Bojnordi MN, Monarch: a durable polymorphic memory for data intensive applications. IEEE Transactions on Computers. 2022.
38. Li H, Lu X. Challenges and trends of big data analytics. In proceedings of IEEE Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing. 2014. pp. 566–567.
39. Ardagna CA, Bellandi V, Ceravolo P, Damiani E, Bezzi M, Hebert C. A model-driven methodology for big data analytics-as-a-service. In proceedings of IEEE International Congress on Big Data (BigData Congress). 2017. p. 105–112.
40. Rialti R, Zollo L, Ferraris A, Alon I. Big data analytics capabilities and performance: evidence from a moderated multi-mediation model. Technol Forecast Soc Chang. 2019;149: 119781.
41. Slavakis K, Giannakis GB, Mateos G. Modeling and optimization for big data analytics:(statistical) learning tools for our era of data deluge. IEEE Signal Process Mag. 2014;31(5):18–31.
42. Poudel M, Sarode RP, Watanobe Y, Mozgovoy M, Bhalla S. A survey of big data archives in time-domain astronomy. Appl Sci. 2022;12(12):6202.
43. Londhe A, Rao PP. Platforms for big data analytics: trend towards hybrid era. In proceedings of IEEE international conference on energy, communication, data analytics and soft computing (ICECDS). 2017. p. 3235–3238.
44. Alsheikh MA, Niyato D, Lin S, Tan HP, Han Z. Mobile big data analytics using deep learning and apache spark. IEEE Network. 2016;30(3):22–9.
45. Gudivada VN, Irfan MT, Fathi E, Rao DL. Cognitive analytics: going beyond big data analytics and machine learning. In Handbook Stat. 2016;35:169–205.
46. Khalajzadeh H, Simmons AJ, Abdelrazek M, Grundy J, Hosking J, He Q. An end-to-end model-based approach to support big data analytics development. J Comput Languages. 2020;58: 100964.
47. Spangenberg N, Roth M, Franczyk B. Evaluating new approaches of big data analytics frameworks. In proceedings of International conference on business information systems. 2015. p. 28–37.

48. Lee I. Big data: dimensions, evolution, impacts, and challenges. Bus Horiz. 2017;60(3):293–303.
49. Elaraby NM, Elmogy M, Barakat S. Deep Learning: Effective tool for big data analytics. International Journal of Computer Science Engineering (IJCSE). 2016;9.
50. Kumar O, Goyal A. Visualization: a novel approach for big data analytics. In proceedings of second international conference on computational intelligence and communication technology (CICT). 2016. p. 121–124.
51. Osman AMS. A novel big data analytics framework for smart cities. Futur Gener Comput Syst. 2019;91:620–33.
52. Kibria MG, Nguyen K, Villardi GP, Zhao O, Ishizu K, Kojima F. Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks. IEEE Access. 2018;6:32328–38.
53. Raut RD, Mangla SK, Narwane VS, Gardas BB, Priyadarshinee P, Narkhede BE. Linking big data analytics and operational sustainability practices for sustainable business management. J Clean Prod. 2019;224:10–24.
54. Lu Y, Xu X. Cloud-based manufacturing equipment and big data analytics to enable on-demand manufacturing services. Robotics Comput-Integr Manuf. 2019;57:92–102.
55. Al-Ali AR, Zualkernan IA, Rashid M, Gupta R, AliKarar M. A smart home energy management system using IoT and big data analytics approach. IEEE Trans Consum Electron. 2017;63(4):426–34.
56. Holmlund M, Van Vaerenbergh Y, Ciuchita R, Ravald A, Sarantopoulos P, Ordenes FV, Zaki M. Customer experience management in the age of big data analytics: a strategic framework. J Bus Res. 2020;116:356–65.
57. Amini S, Gerostathopoulos I, Prehofer C. Big data analytics architecture for real-time traffic control. Models and technologies for intelligent transportation systems (MT-ITS). IEEE. 2017. p. 710–715.
58. Ragini JR, Anand PR, Bhaskar V. Big data analytics for disaster response and recovery through sentiment analysis. Int J Inf Manage. 2018;42:13–24.
59. Rathore MM, Paul A, Hong WH, Seo H, Awan I, Saeed S. Exploiting IoT and big data analytics: defining smart digital city using real-time urban data. Sustain Cities Soc. 2018;40:600–10.
60. Keshk M, Moustafa N, Sitnikova E, Turnbull B. Privacy-preserving big data analytics for cyber-physical systems. Wireless Networks. 2018. p. 1–9
61. Yassine A, Singh S, Hossain MS, Muhammad G. IoT big data analytics for smart homes with fog and cloud computing. Future Gener Comput Syst. 2019;91:563–73.
62. Shorfuzzaman M, Hossain MS, Nazir A, Muhammad G, Alamri A. Harnessing the power of big data analytics in the cloud to support learning analytics in mobile learning environment. Comput Hum Behav. 2019;92:578–88.
63. Darwish TS, Bakar KA. Fog based intelligent transportation big data analytics in the internet of vehicles environment: motivations, architecture, challenges, and critical issues. IEEE. 2018;6:15679–701.
64. Yacchirema DC, Sarabia-Jácome D, Palau CE, Esteve M. A smart system for sleep monitoring by integrating IoT with big data analytics. IEEE. 2018;6:35988–6001.
65. Gohar M, Muzammal M, Rahman AU. SMART TSS: defining transportation system behavior using big data analytics in smart cities. Sustain Cities Soc. 2018;41:114–9.
66. Garg S, Singh A, Kaur K, Aujla GS, Batra S, Kumar N, Obaidat MS. Edge computing-based security framework for big data analytics in VANETs. IEEE Netw. 2019;33(2):72–81.
67. Hadi MS, Lawey AQ, El-Gorashi TE, Elmirghani JM. Patient-centric cellular networks optimization using big data analytics. IEEE. 2019;7:49279–96.
68. Ur Rehman MH, Ahmed E, Yaqoob I, Hashem IAT, Imran M, Ahmad S. Big data analytics in industrial IoT using a concentric computing model. IEEE Commun Mag. 2018;56(2):37–43.
69. Babar M, Arif F, Jan MA, Tan Z, Khan F. Urban data management system: towards big data analytics for internet of things based smart urban environment using customized hadoop. Future Gener Comput Syst. 2019;96:398–409.
70. He QP, Wang J. Statistical process monitoring as a big data analytics tool for smart manufacturing. J Process Control. 2018;67:35–43.
71. Wang K, Li H, Feng Y, Tian G. Big data analytics for system stability evaluation strategy in the energy internet. IEEE Trans Ind Inform. 2017;13(4):1969–78.
72. Chang V. A proposed social network analysis platform for big data analytics. Technol Forecast Soc Chang. 2018;130:57–68.
73. Rajeswari S, Suthendran K, Rajakumar K. A smart agricultural model by integrating IoT, mobile and cloud-based big data analytics. Intelligent Computing and Control (I2C2). IEEE. 2017. p. 1–5
74. Barbeito-Caamaño A, Chalmeta R. Using big data to evaluate corporate social responsibility and sustainable development practices. Corp Soc Responsib Environ Manag. 2020;27(6):2831–48.
75. Bangui H, Ge M, Buhnova B, Hong Trang L. Towards faster big data analytics for anti-jamming applications in vehicular ad-hoc network. Trans Emerg Telecommun Technol. 2021;32(10):e4280.
76. Dankwa-Mullan I, Zhang X, Le PT, Riley WT. Applications of big data science and analytic techniques for health disparities research. The Science of Health Disparities Research. 2021. p. 221–242
77. Nayak J, Kumar PS, Reddy DKK, Naik B, Pelusi D. Machine learning and big data in cyber-physical system: methods, applications and challenges. Cognitive engineering for next generation computing: a practical analytical approach. 2021. p. 49–91
78. Luthra H, Nihith TAS, Pravallika VSS, Shree RR, Chaurasia A, Bansal H. New paradigm in healthcare industry using big data analytics. In IOP conference series: materials science and engineering. 2021. Vol. 1099, No. 1, pp. 012054
79. Zheng X. The application of big data technology in network marketing. J Phys Conf Ser. 2021;1744(4):042200.
80. Ye X. The application of big data in the political and computer education of colleges and universities. J Phys Conf Ser. 2020;1648(3):032104.
81. Moharm K, Eltahan M. The role of big data in improving e-learning transition. In IOP Conference Series: Materials Science and Engineering. 2020; Vol. 885. No. 1. p. 012003.
82. Liang JH. Application of big data technology in product selection on cross-border e-commerce platforms. J Phys Conf Ser. 2020;1601(3):032012.
83. Jin W, Jitao Y. Management innovations research of the logistics enterprises based on the big data environment. J Phys Conf Ser. 2020;1616(1):012010.

84. Ramesh A, Rajkumar S, Livingston LJ. Disaster management in smart cities using IoT and big data. J Phys Conf Ser. 2020;1716(1):012060.
85. Kend M, Nguyen LA. Big data analytics and other emerging technologies: the impact on the Australian audit and assurance profession. Aust Account Rev. 2020;30(4):269–82.
86. Oussous A, Benjelloun FZ, Lahcen AA, Belfkih S. Big data technologies: a survey. J King Saud Univ-Comput Inform Sci. 2018;30(4):431–48.
87. Chen CP, Zhang CY. Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inf Sci. 2014;275:314–47.
88. Agarwal S, Sonbhadra SK, Singh Punn N. Software Testing and Quality Assurance for Data Intensive Applications. In the International conference on evaluation and assessment in Software Engineering 2022. p. 461–462, 2022.
89. Fekete JD. Visual analytics infrastructures: from data management to exploration. Computer. 2013;46(7):22–9.
90. Ahmadvand H, Dargahi T, Foroutan F, Okorie P, Esposito F. Big data processing at the edge with data skew aware resource allocation. In 2021 IEEE conference on network function virtualization and software defined networks (NFV-SDN). 2021. 81–86.
91. Ahmadvand H, Foroutan F, Fathy M. DV-DVFS: merging data variety and DVFS technique to manage the energy consumption of big data processing. J Big Data. 2021;8(1):1–16.
92. Ahmadvand H, Goudarzi M. Significance-aware approach to improve qor of big data processing in case of budget constraint. J Supercomput. 2019;75(9):5760–81.
93. Ahmadvand H, Goudarzi M, Foroutan F. Gapprox: using gallup approach for approximation in big data processing. J Big Data. 2019;6(1):1–24.
94. Ahmadvand H, Goudarzi M. Using data variety for efficient progressive big data processing in warehouse-scale computers. IEEE Comput Architect Lett. 2016;16(2):166–9.
95. Preuveneers D, Joosen W. Automated configuration of NoSQL performance and scalability tactics for data-intensive applications". In Informatics. 2020. Vol. 7, No. 3, p. 29.
96. Tuli S, Sandhu R, Buyya R. Shared data-aware dynamic resource provisioning and task scheduling for data intensive applications on hybrid clouds using aneka. Futur Gener Comput Syst. 2020;106:595–606.
97. Nguyen M, Alesawi S, Li N, Che H, Jiang H. A black-box Fork-join latency prediction model for data-intensive applications. IEEE Trans Parallel Distrib Syst. 2020;31(9):1983–2000.
98. Rao B, Liu Z, Zhang H, Lu S, Wang L. SODA: A semantics-aware optimization framework for data-intensive applications using hybrid program analysis. In 2021 IEEE 14th International Conference On Cloud Computing (CLOUD). 2021. p. 433–444.
99. Rawas S, Zekri A. EEBA: energy-efficient and bandwidth-aware workload allocation method for data-intensive applications in cloud data centers. IAENG Int J Comput Sci. 2021;48:3.
100. Sfiligoi I, Würthwein F, Davila D. Data intensive physics analysis in azure cloud. In: Pasumpon Pandian A, Fernando X, Haoxiang W, editors. Computer networks, big data and IoT. Singapore: Springer Nature Singapore; 2022.

Publisher's Note