An Analysis of Data Processing for Big Data Analytics

[1]Steve Blair and [2]Jon Cotter
[1]Department of Mathematics & Computing, Lander University, Greenwood, SC 29649, United States.
[1]steveblair@lander.edu, [2]cottermail@hotmail.com

Article Info
Journal of Computing and Natural Science (http://anapub.co.ke/journals/jcns/jcns.html)
Doi: https://doi.org/10.53759/181X/JCNS202101019
Received 10 April 2021; Revised form 22 May 2021; Accepted 06 July 2021.
Available online 05 October 2021.

Abstract – The need for high-performance Data Mining (DM) algorithms is being driven by the exponentially increasing data availability such as images, audio and video from a variety of domains, including social networks and the Internet of Things (IoT). Deep learning is an emerging field of pattern recognition and Machine Learning (ML) study right now. It offers computer simulations of numerous nonlinear processing layers of neurons that may be used to learn and interpret data at higher degrees of abstractions. Deep learning models, which may be used in cloud technology and huge computational systems, can inherently capture complex structures of large data sets. Heterogeneousness is one of the most prominent characteristics of large data sets, and Heterogeneous Computing (HC) causes issues with system integration and Advanced Analytics. This article presents HC processing techniques, Big Data Analytics (BDA), large dataset instruments, and some classic ML and DM methodologies. The use of deep learning to Data Analytics is investigated. The benefits of integrating BDA, deep learning, HPC (High Performance Computing), and HC are highlighted. Data Analytics and coping with a wide range of data are discussed.

Keywords – Heterogeneous Computing (HC), Internet of Things (IoT), Big Data Analytics (BDA), Data Mining (DM), Machine Learning (ML).

I. INTRODUCTION

Any data containing a wide range of data kinds and formats is described as Heterogeneous Computing (HC). Due to lacking values, huge dataset redundancies, and inauthenticity, they are perhaps unclear and poor quality. To fulfil corporate information needs, it is challenging to combine diverse data. Internet of Things (IoT) [1], for example, produces HC. The following four characteristics often appear in IoT data: They are, first and foremost, diverse. The obtained data types are HC due to the range of data gathering equipment. Second, they're large. It is necessary to employ and disseminate large data sets acquisition devices. Data must be preserved not just in the present, but also in the past for a set period of time. Finally, time and place are inextricably linked. Data is collected at a predetermined location, with each piece of information being timestamped. Data from the IoT has a lot of time and spatial correlation. Fourth, valuable data makes only a tiny percentage of all huge data. During the gathering and distribution of data in the IoT, a large amount of noise might be recorded. Only a tiny portion of the data collected by collection devices is useful. Data heterogeneousness may be classified as follows:

- Syntactic heterogeneousness exists when two sources of data are not represented in comparable language.
- Conceptual heterogeneousness, commonly known as logical mismatch or semantic heterogeneousness, refers to discrepancies in modeling the same subject of interest.
- The term "terminological heterogeneity" refers to the fact that different data sources use different names for the same thing.
- Semiotic heterogeneity, also known as pragmatic heterogeneity, relates to people's different perceptions of the same object or situation.

There are four layers of data representations. Level 1 is a collection of raw data of various forms and origins. "Unified representations" is the second level. Data that is HC must be unified. Furthermore, excessive data might result in significant cognitive and information processing expenses. Individual qualities are converted into information in the form of "what when where." Aggregation is the third level. Spatial data may be expressed in a natural way using spatial grids with theme properties. Classification, consolidation, and other processing operators are examples. Aggregation facilitates visualization and allows for a simple query. "Condition recognition and representations" is the fourth level. At level 3, the condition at a place is described using temporal descriptions derived by suitable operators. The last step in scenario identification is the classification technique, which uses domain knowledge to classify each cell.

The importance of documentation for future queries cannot be overstated. Explicit schema definitions in Document Type Definition (DTD), XML Scheme Definition (XSD), or Structured Query Language (SQL) may be collected directly from sources and incorporated into a meta-model for relational databases and certain Extensible Markup Language (XML)

files. Data is translated using the XML method. The problematic aspect is semi-structured metadata, which has implicit frameworks (- for example JSON, without XSD, XML, or partly structured CSV or Excel file). As a result, the element "structural metadata discovery" (SMD) is in charge of extracting implicit information (such as entity kinds, relation types, and restraints) from semi-structured datasets. The importance of data handling cannot be overstated. For an accurate assessment of HC huge dataset, detailed metadata is essential. Some reports include metadata, but for research purposes, many additional information, such as those concerning the exact sensor utilized in data gathering, are required. When data is obtained under pressure and in stressful settings, collecting documentation and data authenticity is a huge difficulty.

The issues posed by large datasets, complex, dynamic and scattered data properties are addressed by large data sets algorithms, which focus on algorithm design [2]. The phases of the challenges are as follows: Data fusion methods are used to pre-process HC, fragmented, inconsistent, limited, and multi-source data. Secondly, following pre-processing, dynamic and complicated sets of data are extracted. Thirdly, the global acquired knowledge from local training and modeling fusion is put to the test, and appropriate data is supplied back into the pre-processing step. The models and variables are then tweaked in response to the input. Sharing data is not only a guarantee of smooth progress at each step, but it is also the goal of massive data analytics. This paper has been arranged as follows: Section II presents an analysis of data processing techniques for Heterogeneous Computing (HC) and Big Data Analytics (BDA). Section III presents an analysis of Big Data, Big Data Tools, Big Data Analytics (BDA). Section IV presents an overview of the conventional Machine Learning (ML), Big Data Analytics (BDA), and Data Mining (DM). Lastly, Section V draws the final remarks of research and recommendations.

II. DATA PROCESSINNG TECHNIQUES FOR HETEROGENEOUS COMPUTING (HC) AND (BIG DATA ANALYTICS (BDA)

Data Cleaning

Data cleaning [3] is the procedure of identifying and modifying or deleting data that is incomplete, incorrect, or inappropriate in order to improve data quality. Health data, for instance, has a high level of intricacy and noise due to its multimodal and multisource nature. In particular, there are issues with incomplete data and impurities in large amounts of data. Because data quality affects data quality, which in turn affects decision-making, it's vital to create efficient large data sets cleaning methods to enhance data quality so that effective and appropriate judgments can be made.

A variable with a missing value is one which has not been put into a dataset but has a real value. It's common to utilize simple (non-stochastic) interpolation. Imputed values are used to fill in data that is not full (for example, mean, median, or mode). Instead of providing accurate results when data isn't completely missing, simple imputation does so. Simple imputation is prone to underestimate systematic deviation, skew associations between variables, and provide inaccurate p-values in sampling methods when there is moderate to large quantities of missing data. For the majority of missing data issues, this strategy should be avoided.

We were able to plug in some additional unknown values by studying linear relationships. People may attempt to investigate the relationships between the random parameters and the conventional variables in this situation. By examining more precise correlations, unknown values might be filled in. When dealing with a dataset containing missing values, there are numerous approaches that may be used. The most prevalent are: 1) deleting cases with unknown parameters; 2) filling in missing parameters by examining case similarities; 3) filling in missing parameters by examining variable correlation; and 4) employing tools that can manage these values.

An unimportant property may be present in a database. Relevance evaluation as correlation assessment and selection of variable subsets could be employed in the discovery of qualities, which do not add to the predictions or classifications tasks. Integrating these features could somewhat cause the learning step to slow down and perhaps mislead. Data cleansing and interoperability are often done as part of the preparation process. Redundancies in the final dataset may be caused by variations in variable or dimensional nomenclature. Data cleaning may be used to find and eliminate inefficiencies which may have arisen as a consequence of system integration. The goal of data cleaning and data reduction is typically seen to be the elimination of superfluous data.

Data Integration

Datasets are matched and combined based on common variables and characteristics in system integration or aggregation. We may blend unstructured and structured datasets to elicit fresh insights using advanced statistical data analysis tools. Nevertheless, "clean" data is required. Data fusion methods are used to combine and match disparate information in order to create or improve reality reconstructions that aid DM. Methods for combining conventional and machine-produced datasets at the mid-level work effectively in most cases. High-level data fusion jobs that combine several unstructured analogue sensing data, on the contrary, remains difficult.

Data integration technologies [4] are progressing toward harmonization of unstructured and structured data, and semantics functionalities will be added. Unstructured data must often be unstructured and structured data sources and kinds must be merged into a single data layer. The majority of data integration solutions have a main integration paradigm that is based on relational or XML datasets. Advanced Data Virtualisation Technologies have been presented that employ an expanded integrative database schema with the capacity to read/write any sorts of data in its native format, including relational, multivariate, semantics metadata, hierarchical, and indexed files, among others.

It is difficult to bring together disparate data streams. One of the grounds for this is because there are frequently no unique IDs between entries from two separate databases. It's possible that the factors that need to be incorporated aren't immediately apparent. With HC, the importance of data seems to be discovered iteratively, and the most beneficial data may be methodically blended. In order to deal with data heterogeneity, the following interface was proposed: One of the most important aspects of a Business-to-Business (B2B) system is the integration of models. The "catalog mirroring issue" refers to identifying relationship between catalogue items.

Query-centric and federated techniques to data integration, as well as knowledge extraction from independent, morphologically diverse, and remote data sources, are of particular interest. 5. The following methodologies may be used to integrate unstructured and structured data:

- Natural language processing (NLP) pipeline [5]: Unstructured data may be processed directly using natural language processing (NLP).
- Recognizing and connecting entities: A crucial step is to retrieve structured datasets from unstructured datasets. Dataset extraction processes such as entity identification, connection extraction, and ontological retrieval may help address part of the issue. These technologies assist in the automated creation of semi-structured information. Certain forms of data extraction issues have developed frameworks, but their adoption is still confined to early adopters.
- Integrating unstructured and structured data with open data: Organizations in open data may be used to recognize entities (individuals, companies, and locations), which can then be used to classify and arrange text contents. Unstructured and structured data may be linked using named entity identification and linking technologies like DBpedia Spotlights.

Data input mistakes, data structure incompatibility, and semantic incongruities in company operating descriptions are three types of data problems when integrating data from disparate technologies together. Data Warehouses (DW) and ETL (Extract, Transform, and Load) have historically been used for system integration (ETL). An alternative data integration method called "Data Virtualization (DV)" [6] has been more popular in recent years. It's a "Composite Database" that uses the term "Data Virtualization." It is possible to reduce data integration costs and timeliness by standardizing corporate data and transforming it.

Unlike DW, DV uses logical views to describe data cleansing, data joining, and transformations automatically. By enabling for the concatenation of logical views, DV provides for flexibility and reusing. Data type incompatibilities and semantics incompatibility in data are usually avoided with enterprise information standardizing. DV is not a substitute for DW. But it may be utilized to delegate some of DW's analytical tasks to a third party. Multidimensional data structures and large-scale data processing all need the use of DW.

As businesses become more exposed to cellular and cloud-based apps, as well as the sensor-driven IoT, data warehouses are developing as a strong solution to the issues of data integration (IoT). Datasets are reservoirs for enormous amounts and types of organized and unstructured data. Data lakes are better suited to processing less-structured data. Nevertheless, obstacles with data lake integrated solutions encompass, but are not restricted to: 1) enhanced metadata governance over raw dataset obtained from datasets; 2) interacting with systemic metadata from datasets; and 3) categorising data and metadata with contextual data to prevent misunderstandings. Dumping all information into a digital lake with no metadata or data governance would result in an "information swamp," in which the data lake is scarcely useable since the architecture and interpretations of the information are unknown.

Data Normalization and Dimension Reduction

The data dimensionality [7] should be reduced for numerous reasons. High-dimensional data, for starters, presents computing difficulties. Second, in certain cases, a high dimensionality training algorithm may have weak generalization skills (As an example, in closest neighbor classifiers, the sample complexity grows exponentially with the dimension). Lastly, feature reduction may be utilized to determine the data's meaningful architecture, understandability, and visualization.

In ML and DM, Feature Subset Selection (FSS) [8] is a common activity. Feature selection challenges may be approached using a variety of approaches, including traditional Simulated Annealing, Hill Climbing, and Genetic Algorithms. An FSS technique may bring numerous benefits in terms of dimensionality reduction: 1) a more rapid initiation of the final categorization algorithm, 2) an increase in the understandability of the ultimate classifying model, and 3) an increase in classification reliability Feature rankings and subgroup selection are two ways to feature evaluation. In the first method, characteristics are sorted according to a set of criteria, and then those that exceed a certain threshold are chosen. The optimum feature subset is found by searching a collection of feature subsets. Furthermore, the second strategy may be broken down into three components: 1) Filter strategies: users first choose the characteristics, then use that subset to operate a classification algorithm; 2) Embedded approaches: features selection is done as component of a classification technique; and 3) wrapper strategies: a classification technique is used to find the best features in a dataset.

There is often a huge overlap in the data processed by a group of variables in databases with a high set of variables. Checking a correlation matrix derived from correlation test is one straightforward technique to find redundancy. A way for reducing complexity is factor modeling. It's helpful for figuring out why a bunch of variables are correlated. Factor analytics is a technique for reducing the quantity of parameters and detecting structure in the interactions between them.

As a result, Factor Analyses is frequently used to find structures or reduce data. PCA is also effective whenever there is information on a number of variables with some redundancies. In this case, redundancy refers to the fact that some of parameters are linked. PCA is a popular and commonly used technique that is quick, effective, and easy. PCA may aid in a variety of ways, including:

- Learning sophisticated models from high-dimensional data takes a long time and may lead to classifier. In most models, the number of variables grows exponentially as the number of dimensionalities increases. PCA may also be used to whiten the representations, which restructures the proportions of the data and, in certain situations, increases accuracy.
- PCA understands a depiction, such as a distribution function for new information, that is also used as a whole model.
- PCA compresses data by substituting it with a low-dimensional abstraction.

PCA or EFA (Exploratory Factor Analysis) involves the following critical phases: Data preparation, such as incorrect value filtering; 2) In the case of an EFA approach, pick a factor model and a factoring method (for example, maximum probability) based on whether EFA (uncovering latent structure) or PCA (data reduction) is more suited to the study's goals; 3) Calculate the number of elements and factors to be extracted; and 4) retrieve the elements and factors. When it comes to determining how many elements to keep in a PCA, there are numerous parameters to consider. These incorporate: 1) evaluating the components centred on previous hypotheses and experiences; 2) determining the elements considered necessary to compensate for a certain amount of accumulated variation in the dependent variable (for instance, 80 percent); and 3) defining the number of elements to preserve based on the eigenvector of the covariance matrix among some of the factors.

Before certain methods to be used successfully, the data must be normalized (standardized). It is through this process of normalization that each unique variable's variation is reduced to one unit (or standardization). This normalization (standardization) has the effect of equating the relevance of all variables in terms of their variability. Before running the PCA, the data is generally standardized. Section III presents an analysis of Big Data, Big Data Tools, Big Data Analytics (BDA).

III. BIG DATA, BIG DATA TOOLS, BIG DATA ANALYTICS (BDA)

In the field of large data sets, we analyze, extract, and otherwise deal with complex and big data for standard data-processing commercial applications. The statistical significance of data with a large number of fields (columns) is higher, but the false alarm rate is higher for data with a large number of features or columns. Managing massive datasets has a number of challenges, including the need for effective data collection and storage as well as analysis and retrieval. Volume, variety, and velocity were the first three concepts to be associated with enormous data collections. Only observations and samples were previously allowed because of sampling concerns with huge data analysis. Due of the sheer volume of information in enormous data sets, standard software cannot manage it in a timely or cost-effective manner.

It is currently more common to use the term "Big Data" to describe the utilize of advanced data analytics methods [9] to extract value from enormous datasets rather than a particular dataset size. Even while this new information environment has a vast quantity of data, it isn't the most significant aspect of it. "Study of data sets may be used to discover new connections in order to "identify latest developments, prevent illnesses, and fight crime, among other things." In sectors such as Online searches, finance, medical informatics, spatial analysis, urban computing, and corporate information systems, academicians, company executives, healthcare professionals, marketing, and authorities all face challenges with enormous data sets. Meteorological, genetics, connectomics, sophisticated physics models, physiology, and scientific research all have limits that scientists face.

As information is recorded by technologies such as smartphones gadgets, cheap and innumerable data-sensing IoT devices, aerial (satellite imagery), application logs, camera systems, recorders, radio-frequency identification (RFID) readers, and sensor networks, the size and amount of available huge dataset has expanded dramatically. Since the 1980s, the globe's technology per-capita ability to store data has steadily increased each 40 months; in 2012, 2.5 exabytes (2.5260 bytes) of data were created per day. As pert DC research, worldwide data volume is predicted to increase rapidly from 4 – 44 zettabytes from 2013 – 2020. In reference to IDC, there might be 163 zettabytes in 2025. One issue that huge companies have is choosing who should be in charge of large-data efforts that touch the whole company.

Big data [10] may be challenging to handle and analyze using database software and desktops analytical software products utilized to show data. "Parallel processing programs operating on dozens, thousands, or even millions of computers" may be required for large data sets management and analysis. What constitutes "large data sets" varies according to the skills of individuals who analyze it and the technologies they use. Additionally, as technology advances, huge dataset becomes a changing target. Whenever associated with multiple terabytes in data at a first instance, some firms might require to re-evaluate dataset management strategies. For others, hundreds or thousands of terabytes may be required before size of data would become a serious concern.

In Continuous Auditing (CA) systems [11], there are inconsistencies between large amounts of data and current data analytic applications. As the first three Vs (Volume, Velocity, and Variety) are established in a CA system, a consistent data flow, data recognition, and data collection gaps emerge. Table 1 lists the issues that each of the gaps brings. Data authentication pertains to recordings that connect two or more pieces of data about the identical person or organization that

were previously recorded independently. Identification is simple when data is organized. In a massive data audit, however, when most of the data is unstructured, authentication becomes more challenging. Data manipulation and inadequate data are the most common causes of data integrity issues.

The most essential challenge for the CA of big data is data integrity, which refers to dependent data throughout programs and throughout an enterprise. Confidentiality refers to the fact that specific data or relationships between data points are secret and cannot be shared with others. Data may be readily connected with other data in the age of large data sets. Once sensitive information has been exposed, it may spread quickly and link to a significant number of connected information. As a result, large data sets secrecy becomes even more critical in order to protect brand reputation and gain a competitive edge. Data aggregation is required for the regular functioning of auditing with large data sets in order to properly consolidate and reduce the huge data, which is likely to come from several sources (see Table 1).

Table 1: Big Data Analytics (BDA) in Continuous Auditing (CA)

BDA in CA	
The 4Vs of Big Data Analytics (BDA)	Volume, Variety, Veracity, Velocity
Gaps in Big Data (BD)	Data identification, Data consistency, Data aggregation, Data integrity, Data confidentiality
Continuous Auditing (CA) challenges	Contradictory datasets, Imperfect datasets, Dataset with different identity, Dataset with various setups, Asynchronous datasets, Illegitimately tempered datasets, searching encrypted datasets, auditing aggregated datasets, Auditing encrypted datasets.

Heterogeneousness is among the most fundmaental facets of big data [12]. By default, data from a range of sources comes in a variety of types and representation formats, and it may be combined, correlated, and shown in a variety of ways. Because of the diversity of large data sets, it's necessary to gather and analyse structured, semistructured, and even entirely unstructured data all at once. Large-scale, complex datasets generated from various sources are referred to as "large data sets." When evaluating large amounts of data, gathering and storing mixed data based on several patterns or criteria might be problematic (HC mixture data). It's crucial to be aware of your surroundings. Current production data, for example, is irrelevant in the setting of users' histories, plans, habits, activities, and locations. In the context of large data sets, contextualisation is a potential approach for merging different data streams to enhance the effectiveness of a mining project or classifier. Even more importantly, the use of awareness " has been found to reduce the consumption of resources by concentrating big data creation operations on just the sources that are expected as most likely based on the current (and sometimes application-specific) context. There are several inconsistencies with large data sets:

- Large data sets aims to recognise, but it also poses a danger to do so. The “identity dilemma” is a word describing a scenario in which two persons are confused about their identities.
- The reliance on tiny data inputs in BDA generates a transparency dilemma. Limited data sources are combined to build huge datasets. This data is gathered in a non-obtrusive manner. Large data sets purports to make the whole world more accessible, but the data is gathered discreetly, as well as the tools and methods used to evaluate it are buried behind levels of geographical, legal, and mechanical secrecy by design.
- Data sensors and pools, rather than the hands of regular citizens, hold the majority of power paradoxically. It is impossible to achieve aspirational objectives in massive data sets unless privacy, autonomy, openness and identity safeguards are included in from the beginning.

The timeliness of the data affects the hidden patterns in large datasets. Consequently, it is necessary to build a critical notion connected to analytical value so that we can choose which data should be removed and which data should be maintained. To build a BDA system, you'll need to do the following:

- Data loading [13]: To load data from a number of sources, software must be developed. A system must cope with Hadoop's distributed nature on the one hand, while the source of data must be non-distributed on the other. The system must be able to monitor and manage corrupted records.
- Data parsing [14]: Prior to using Hadoop, the majority of data sources give data in a certain format. . Some formats, such as JSON, are difficult to interpret since a record might include many lines of text rather than just one.

A BDA system must allow for fast iterations in order to accurately examine the data.

1) Descriptive analytics, which involves describing and summarizing patterns of knowledge; 2) predictive analytics, which integrates statistical and forecasting modeling to survey futuristic probabilities; and 3) prescriptive analytics, which aids analysts in the process of making decision by identifying actions and evaluating the effects.

Many different types of databases have been used to examine enormous data sets: Massively Parallel Processing (MPP), distributed systems, and in-memory or non-relational databases. MPP databases are known for their quick query response times and flexibility in terms of platform scaling. Not NoSQL (non-relational) databases are the only ones that can process and store unstructured data, which is why they're so popular for web and mobile app development and deployment. Due to the lack of disk I/O and the ability to respond instantly, in-memory databases are a great alternative to

traditional relational databases. It has also been shown that in-memory databases may speed up the retrieval and scoring of analytical models in advanced BDA. Solutions for large data sets include, but are not limited to, MarkLogic, RapidMiner, Talend, Tableau, Splunk, and Hive. With Hive, even the largest distributed databases are easy to manage and query. It is an analytics firm that specializes in obtaining machine datasets from several data sources, e.g., websites and sensors. Tableau is an application for creating scatter plots, maps, and other data visualizations. It is possible to build, test, and deploy data management and integration applications using Talend, which is available for free and open source use by anybody.. It's a Graphical User Interface (GUI) for developing, maintaining, and giving projected insights. Whenever handling big data, MarkLogic is able to provide its consumers with real-time updates and notifications.

When velocity hits a critical point, large data sets approaches will almost probably be the only way to stay up. This method is especially beneficial for frequently changing data, since it allows data to be extracted and analyzed fast from massive data sets. Computer complexity may be reduced by using massive data sets technology, which distributes difficult processing tasks over a large number of nodes. The foundation is made up of Hadoop and HDFS. Oozie, Zookeeper, Elastic MapReduce (EMR), and Flume are some of the most popular systems for managing large data sets. Both numerical and nominal data may be processed using MapReduce. It is capable of completing a significant task in a short period of time. Algorithms, on the other hand, must be changed, requiring a thorough understanding of systems engineering.

YARN has made it possible for Hadoop [15] to handle a broad variety of programming paradigms, including batch and near-real-time outputs. Many high-quality software solutions are available to help you take use of the advantages of massive data collections. For example, Dell's Kitenga Analytics Suite is a market-leading huge data sets searches and data analytics platforms, which were developed to mix numerous kinds of data into readily deployable visualizations. Kitenga supports Hadoop, allowing for the incorporation of distinct dataset sources and costlier storage of rising datasets amounts. It's possible for Kitenga to examine Hadoop discoveries and index the data into a searchable form with integrated visualization capabilities utilizing information visualization tools that connect directly to HDFS files. Big data computing system managers commonly use the following three methods:

- Within the company's cluster of computers Hadoop clusters may be built and maintained by networked computers in the organization's internal network for long-term storage of sensitive and unique dataset.
- External compute cluster In the IT business, outsourcing infrastructure parts to "utility computing" service providers is growing increasingly prevalent. Renting pre-built Apache Hadoop systems and digital storage solutions is made simple for system administrators by a number of firms.
- A hybrid computing cluster Using services to provide external compute cluster capabilities for on-demand large data sets analysis workloads while simultaneously constructing a typical internalized computing clusters for long-range storage of data is a common hybrid solution.

As shown in Table 2, the analysis of huge datasetsets is broken down into many phases, each with its own set of challenges. After the analytical process, there are still a lot of challenges to deal with. For example, large data sets must be handled in context, which may be noisy, varied, and devoid of a predictive model. As a consequence, keeping track of authenticity and dealing with errors and uncertainties is increasingly vital. In addition to the normal technical hurdles of large data sets, there are other concerns to consider. 1) making data accessible by integrating meta-data and permitting the incorporation of disparate dataset silos; 2) mitigating regulatory issues that relate to the privacy and ownership of data; and 3) optimizing the value of existing available open data and linked data sources. To mention a few, Large data sets's issues include scalability, heterogeneousness, a lack of structured connectivity, quality of data, confidentiality, and security. Data management, analytics, and data intelligence must be done deliberately and comprehensively to fulfill large data sets's full potential.

Table 2: Pipeline and BDA challenges

Aspect	Description
Key phases in Big Data (BD) analysis	Interpretations,Modeling/analysis, Representation/aggregation/integration, Annotation/extraction/cleaning, Recording/recording
Problems during the phases	Human collaborations, privacy, scale, timeliness, heterogeneity

For even the most established of links, Data Analytics [16] can illustrate how it may fall apart or how a new pattern may develop in its place. The availability of enormous datasets does not eliminate typical statistical traps, such as sampling error and sample bias, according to one criticism of BDA. Claims about the widespread availability of sensing devices and other large data sets sources are often inflated or disregard particular and regular causes of bias. Future needs for large data sets technologies that should be addressed by the next generation.

- Support Internet expansion – As more people use the internet, large data sets technology will be required to manage increasing amounts of data.
- Real-time processing – Traditionally, Large data sets analysis was done in batches using historical data. Stream processing technologies like Apache Storm have been more widely accessible in recent years, allowing for new application possibilities.

- Handle sophisticated data structures – Large data sets technologies must be able to handle data structures such as graphs and maybe other forms of more intricate data structures.
- Effective indexing — Indexing is vital in handling large collection of records and their related metadata since it is required for online data search.
- Multi-server and cloud flexible choreography of services – It's difficult to keep data consistent across multiple data stores with the platforms we have today because they weren't designed for the cloud.
- Simultaneous data processing – The ability to handle large numbers of users at the same time requires the ability to process massive amounts of data continuously.

Section IV presents an overview of the conventional Machine Learning (ML), Big Data Analytics (BDA), and Data Mining (DM).

IV. CONVENTIONAL MACHINE LEARNING (ML), DATA MINING (DM), BIG DATA ANALYTICS (BDA)

Finding outliers is one of Data Mining (DM) tasks [17]. The four kinds of computer-based outlier detection approaches include statistical methodology, density-based outlier technique, distance-based strategy, and deviation-based strategy. The LOF (Local Outlier Factor) is a density-based approach to spotting local outliers. LOF is used to compare a location's density to the density of its neighbors. If the former is much lower than the latter, it suggests that the point is an outlier since it is in a smaller percentage zone than its neighbors (and the LOF value is larger than one). LOF's main drawback is that it only works with numerical quantities. Clustering is a different approach to spotting outliers. It is possible to utilize clustering techniques to identify groups of data that are geographically dissimilar. Outliers are features that do not fit into any categories once data has been clustered.

Numerous issues arise when dealing with non-numerical variables, such as alpha parameters. Among the algorithms that can handle alpha values in their alpha state is the decision tree algorithm. Only a numerical representation of alpha may be used by other techniques, such as neurons. There are a wide variety of database formats, therefore some databases may include complicated data elements, such as time series or geographic data. To expect a single provider to collect all types of data is impractical. As a consequence, diverse DM approaches for various kinds of data are likely to emerge. "Privacy-preserving data analysis" is a research path aimed at bridging the gap between large data sets protection. "Collective Data Mining (CDM)" is a methodology for dealing with the problem of learning from disparate data.

High-performance data retrieval [18] is often necessary due to the sheer size or complexity of large datasets. High-speed DM necessitates the use of concurrent database systems and additional CPUs. Parallelism's primary purpose is to improve performance. Increase in productivity may be assessed in two ways. Capacity, or how many jobs can be accomplished in a given amount of time, is the first. Responding time, or how long a single piece of work will take to finish, is also an important issue. There are two common metrics used to quantify the scale-up and speed of the two measures. The Large data sets Reference Architecture has been provided and explained, which involves: 1) an information management strategy which consolidates all types of information, such as unstructured data, semi-structured data and structured data; 2) real and packet data feeds; and 3) high-performance in-memory and in-database actionable insights.

A large quantity of unstructured data is used in deep learning algorithms to derive complicated representations. Deep learning systems allow for global and non-local generalization. It builds abstractions from unsupervised datasets without requiring the intervention of humans. As a result, deep learning is a viable method for data analytics when the raw data is primarily unlabeled and unsorted [19]. Collaborating on huge datasets using High Performance Computing (HPC), High-throughput Computing (HC), Deep Learning (DL) enhances computer intelligence and performance.

Table 3: Deep Learning (DL) and Traditional Machine Learning (ML) and Data Mining (DM) Employed in Healthcare

Algorithms and ML/DL categories	Advantages	Disadvantages	Healthcare samples
Density-oriented Clustering	Handles both complex and non-static datasets, determine outliers and random shapes	Tricky, slow parametric selection, not well for big data	Biomedical clustering of images identifying network biclique
Partitioning Clustering	Fast, simple, and useful in dealing with big data	High sensitivity to outliers, noise and initialization	Depression clusters, risks of re-admission forecasting
Hierarchical Clustering (HC)	Capacity of visualizaiton	Low, slow in accuracy, poor-quality of virtualization of big data, utilizing big data of memory	Micro-array clustering of data, grouping of patients centered on the length of stay in the clinic.
Support Vector Machines (SVMs)	High accuracies	Slower training, computationally costlier	Children healthcare, image-oriented MR forecasting
Decision Trees (DTs)	Easy and simple to implement	Overfitting, space limitations	Brain MRI categorization, clinical forecasting

Nural Networks (NNs)	Handle noisy datasets, identify non-linear connections	Low, slow accuracies, black-box frameworks, computationally costlier	Blood sugar level forecasting, cancer, heart rate variabilities and identification
Ensemble	Generalization, predictive, overcoming overfitting, high performance	Hard to evaluate, computationally costlier	Rate of morality and forecasting, Alzheimer categorization, drug treatment response forecasting
Deep Learning (D)	Bog data, deep architectures, multi-task, unsupervised feature training, semi-supervised learning, generalization	Interpretation difficulty, computationally costlier	Alzheimer diagnostics, MR registration of brain images, clinical decision-making

V. CONCLUSION AND FUTURE RESEARCH

To increase data quality, it's critical to create effective large-scale data cleaning methods. Data lakes and Data virtualization are effective data integration techniques. When reducing data dimensions, EFA and PCA are often utilized. When it comes to dealing with a wide variety of data, heterogeneousness is also a component of the equation. Each step of data analytics has its own set of problems. Real-time computing, dealing with complicated data kinds, and simultaneous data analysis are just a few examples. Data analytics has constraints for conventional ML and DM methodologies. Deep learning can analyze and learn from vast volumes of unsupervised data; hence it has the capability to be utilized in large data sets and analytics where the raw data is mostly unlabeled and unsorted. For HC large data sets, the intersections of Heterogeneous Computing (HC) and Data Analytics, High-Performance Computing (HPC), and deep learning might be a study area. Deep learning methods are used to evaluate data by the first learning high-level data representations pattern using a portion of the available data corpus, then retrieving data abstraction and representation with the remainder input corpus and learnt patterns. The amount of data required for deep learning models to create relevant or great datasets is a topic worth investigating. Future deep learning study in data analytics might focus on this. When it comes to smart healthcare, deep learning has benefits and downsides over regular machine learning (ML). Data quantities, high-dimensionality, unsorted and unstructured datasets, and other challenges have stifled traditional ML and DM approaches. As a result, when it pertains to the BDA, they are limited.

References

[1]. M. V. Ngo, T. Luo, and T. Q. S. Quek, "Adaptive anomaly detection for internet of things in hierarchical edge computing: A contextual-bandit approach," ACM Trans. Internet Things, vol. 3, no. 1, pp. 1–23, 2022.

[2]. Ramesh R., Udayakumar E., Srihari K., and Sunil Pathak P., "An innovative approach to solve healthcare issues using big data image analytics," Int. j. big data anal. healthc., vol. 6, no. 1, pp. 15–25, 2021.

[3]. S. Song, F. Gao, A. Zhang, J. Wang, and P. S. Yu, "Stream data cleaning under speed and acceleration constraints," ACM trans. database syst., vol. 46, no. 3, pp. 1–44, 2021.

[4]. The Mouse Phenotype Database Integration Consortium, "Integration of mouse phenome data resources," Mamm. Genome, vol. 18, no. 11, pp. 815–815, 2007.

[5]. Y. Li, X. Yang, M. Zuo, Q. Jin, H. Li, and Q. Cao, "Deep structured learning for natural language processing," ACM trans. Asian low-resour. lang. inf. process., vol. 20, no. 3, pp. 1–14, 2021.

[6]. A. J. Elmore, C. Curino, D. Agrawal, and A. El Abbadi, "Towards database virtualization for database as a service," Proceedings VLDB Endowment, vol. 6, no. 11, pp. 1194–1195, 2013.

[7]. D. Bera, R. Pratap, and B. D. Verma, "Dimensionality Reduction for Categorical Data," IEEE Trans. Knowl. Data Eng., pp. 1–1, 2021.

[8]. E. Civitelli, M. Lapucci, F. Schoen, and A. Sortino, "An effective procedure for feature subset selection in logistic regression based on information criteria," Comput. Optim. Appl., vol. 80, no. 1, pp. 1–32, 2021.

[9]. S. Kusal, S. Patil, K. Kotecha, R. Aluvalu, and V. Varadarajan, "AI based emotion detection for textual big data: Techniques and contribution," Big Data Cogn. Comput., vol. 5, no. 3, p. 43, 2021.

[10]. A. Arif, T. A. Alghamdi, Z. A. Khan, and N. Javaid, "Towards efficient energy utilization using big data analytics in smart cities for electricity theft detection," Big Data Res., vol. 27, no. 100285, p. 100285, 2022.

[11]. T. Sun, M. Alles, and M. A. Vasarhelyi, "Adopting continuous auditing: A cross-sectional comparison between China and the United States," Manag. Audit. J., vol. 30, no. 2, pp. 176–204, 2015.

[12]. A. N. Henderson, S. K. Kauwe, and T. D. Sparks, "Benchmark datasets incorporating diverse tasks, sample sizes, material systems, and data heterogeneity for materials informatics," Data Brief, vol. 37, no. 107262, p. 107262, 2021.

[13]. S. Goutianos, "Fracture resistance dataset of composites under mixed-mode non-proportional loading," Data Brief, vol. 39, no. 107668, p. 107668, 2021.

[14]. M. Damonte and E. Monti, "One semantic parser to parse them all: Sequence to sequence Multi-Task Learning on semantic parsing datasets," arXiv [cs.CL], 2021.

[15]. Y. Yao, H. Gao, J. Wang, B. Sheng, and N. Mi, "New scheduling algorithms for improving performance and resource utilization in Hadoop YARN clusters," IEEE trans. cloud comput., vol. 9, no. 3, pp. 1158–1171, 2021.

[16]. "Call for papers: Special issue on unlocking genetic diseases by integrating machine learning techniques and medical data," Big Data Min. Anal., vol. 4, no. 3, pp. 221–221, 2021.

[17]. F. Stahl and I. Jordanov, "An overview of the use of neural networks for data mining tasks: Use of neural networks for data mining tasks," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 2, no. 3, pp. 193–208, 2012.

[18]. X. Li, B. Yu, G. Feng, H. Wang, and W. Chen, "LotusSQL: SQL engine for high-performance big data systems," Big Data Min. Anal., vol. 4, no. 4, pp. 252–265, 2021.
[19]. F. Harrou, A. Dairi, F. Kadri, and Y. Sun, "Effective forecasting of key features in hospital emergency department: Hybrid deep learning-driven methods," Machine Learning with Applications, vol. 7, no. 100200, p. 100200, 2022.