International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/IJSRCSEIT

An in-Depth Review of Big Data Analytic Models for Clustering Operations

Trishali Dhote, Prof. Pragati Patil

Department of Computer Science and Engineering, Tulsiramji Gaikwad-Patil of Engineering &Technology, Nagpur, Maharashtra, India

ARTICLEINFO

Article History:

Accepted: 01 Sep 2023

Published: 08 Sep 2023

Publication Issue

Volume 9, Issue 5

September-October-2023

Page Number

30-47

ABSTRACT

The ever-increasing volume, pace, and variety of data in the present environment of data-driven decision-making need creative ways for extracting insightful information sets. A key component of data analysis, clustering techniques are essential for identifying latent patterns and structures in huge datasets. This work conducts a thorough investigation of big data analytic models for clustering operations in response to the pressing requirement to harness the power of big data analytics for effective and accurate clustering process. The necessity for this effort derives from the expanding levels of data volume and complexity that characterise modern information ecosystems. While effective in smaller datasets, conventional clustering approaches fall short when faced with the enormous datasets typical in contemporary applications. As a result, choosing and using the right big data analytic models for clustering have become crucial tasks for both researchers and practitioners. The review procedure used here is defined by a thorough and comprehensive approach. The first stage includes a thorough literature review in which a wide range of big data analytical models are methodically developed. These models cover a broad range of strategies, from hierarchical and model-based approaches to density-based and partitioning techniques. The foundation for the future research is laid by this thorough assessment, which focuses on a detailed analysis of the characteristics and performance measures of each model process. The empirical assessment considers a wide range of factors, including accuracy, computational complexity, scalability, and applicability for various application areas. A comprehensive knowledge of each model's potential and constraints is revealed by closely examining each model's performance across these aspects. This not only encourages a thorough understanding of the models' capabilities but also equips

practitioners with the knowledge they need to carefully choose the best model for their unique clustering jobs.

Keywords: Big Data Analytics, Clustering Operations, Data Complexity, Analytic Models, Empirical Evaluation Process

I. INTRODUCTION

In the era of information abundance, the transformative potential of data analysis has become an irrefutable driver of progress across domains. As the dimensions of data continue to expand exponentially, a fundamental challenge surfaces—how to extract meaningful insights from vast and complex datasets. Clustering, a quintessential exploratory data analysis technique, assumes a paramount role in unveiling latent patterns, structures, and associations within diverse data repositories. In the context of this burgeoning data landscape, the efficacious application of clustering operations has evolved into a pivotal pursuit, and it is this pursuit that catalyzed the inception of the present paper—an in-depth review of big data analytic models for clustering operations.

The compelling need for this work emanates from the evolving contours of data ecosystems, characterized by an unrelenting surge in data volume, velocity, and variety. Traditional clustering approaches, inherently tailored to relatively modest datasets, find themselves outpaced and outperformed by the voluminous datasets that characterize contemporary information landscapes. The inherent limitations of classical clustering techniques, such as scalability constraints and reduced efficacy with high-dimensional and heterogeneous data, necessitate an exploration of new vistas. It is within this context that big data analytic models emerge as beacons of promise, illuminating pathways for more effective, efficient, and accurate clustering in the face of unprecedented data intricacies.

The principal rationale for this study is rooted in the insufficiency of conventional clustering techniques in addressing the demands of modern data scenarios. Big data, while harboring immense potential for insights, presents formidable computational and algorithmic challenges. To bridge this chasm between burgeoning data challenges and effective clustering operations, researchers and practitioners are impelled to navigate the burgeoning landscape of big data analytic models. The primary objective of this paper, therefore, is to provide an encompassing review of these models, elucidating their underlying principles, strengths, limitations, and real-world applicability levels.

The methodological architecture of this endeavor is characterized by a comprehensive two-tiered process. The preliminary phase encompasses an exhaustive review of the extant literature, wherein a diverse and comprehensive array of big data analytic models tailored for clustering operations are systematically collated. This compendium encapsulates a spectrum of methodologies, encompassing density-based, partitioning, hierarchical, and model-based approaches. This thorough review establishes the foundational underpinnings for the subsequent empirical analysis.

The empirical analysis, infused with quantitative rigor, embodies the heart of this study. It involves an intricate evaluation of each identified big data analytic model across a manifold of criteria. These criteria traverse a multidimensional terrain, including efficacy

in pattern extraction, scalability to accommodate vast datasets, computational efficiency, adaptability to diverse data types, and applicability to a variety of domains. Through this intricate web of assessment, a comprehensive understanding of each model's capabilities, limitations, and contextual suitability is meticulously cultivated for different scenarios.

The forthcoming sections of this paper are meticulously organized to facilitate a coherent and progressive presentation of insights. Section II delves into an expansive literature review, spotlighting the diverse gamut of big data analytic models for clustering operations. This contextual foundation segues into Section III, which lays bare the methodological framework underpinning the empirical analysis. Section IV unfurls the empirical findings, delineating the performance and attributes of each model across the spectrum of evaluation criteria. Section V engages in a scholarly discourse, dissecting the implications, practical ramifications, and potential trajectories illuminated by the empirical exploration. The paper culminates in Section VI, encapsulating the collective insights into conclusive remarks that not only emphasize the significance of big data analytic models for clustering operations but also offer guidance for their informed selection and applications.

II. Literature Review

Numerous techniques have been proposed by academics to improve the effectiveness of the clustering process for massive datasets. For instance, a study conducted by researchers [1] examined the complex task of clustering a very large and widely dispersed dataset of at least 100 terabytes. Utilizing a representative subset of the massive dataset to derive an approximation of the result is a prevalent method for addressing this specific challenge. This method provides an estimate of the genuine outcome generated by the entire dataset. Rather than relying on a single random sample, this work employs an ensemble approach to predict the true outcome of a large dataset by utilizing multiple random samples. Our research introduces a novel approach to distributed computing that permits the computation of ensemble outcomes. The proposed method employs random sample data units that are processed within a distributed file system in order to represent a massive dataset as an RSP data model. Individual RSP data blocks are clustered in parallel on the cluster nodes to produce component clustering results. This information is then used to compute the ensemble clustering result. The master node computes the ensemble result after receiving the component results. Given the non-continuous nature of the random samples, which inhibits the use of traditional consensus methods, we propose two novel strategies for incorporating the component clustering outcomes into the collective ensemble outcome. The first methodology entails the formation of a network composed of cluster centers. Afterwards, the graph is partitioned into subgraphs using the METIS algorithm, which helps to identify a set of potential cluster centers. Using a hierarchical clustering technique, the final collection of k cluster centers is then produced. Using the clustering-by-passing-messages procedure of the second methodology, the ultimate aggregation of k cluster centers is generated. The entire dataset was ultimately divided into k clusters using the k-means algorithm. Both synthetic and real-world datasets were utilized in experiments. The results indicate that the use of distributed computing architecture for clustering large datasets and samples is efficient and flexible. In addition, the innovative ensemble clustering techniques demonstrated superior performance compared to the comparison methods.

Computer clusters use data-parallel computation technologies such as Hadoop and Spark for big data analytics, according to research published in [2]. It is common practice for multiple individuals to share a single computer cluster. Multiple occupations may place a substantial burden on one's schedule. The Shortest Job First (SJF) scheduling method is widely acknowledged as the most effective strategy for

reducing the average completion time of tasks. The Shortest Job First (SJF) method, on the other hand, exhibits suboptimal system throughput in situations where a small number of brief activities consume a substantial quantity of resources. This particular characteristic contributes to an increase in the average time required to complete a task. DJSF is proposed as an optimized heuristic task scheduling method. The goal of the DJSF methodology is to reduce the average Job Completion Time (JCT) and increase system throughput by strategically scheduling activities to maximize the number of completed tasks within a given time frame. Our research entails performing exhaustive simulations with Google cluster data. The DJSF methodology increases system throughput by 42.19 percent and reduces the average Job Completion Time (JCT) by 23.19 percent in comparison to the SJF approach. The application of the work packaging technique results in a 55.4% increase in job completion efficacy compared to the Tetris method. This enhancement enables computer systems to coordinate a greater number of tasks in a shorter amount of time.

In the disciplines of data mining and pattern recognition, the possibilistic c-means method (PCM) is a prominent fuzzy clustering technique, according to research published in [3]. It has found extensive use in image analysis and information discovery. However, PCM's primary purpose is to manage tiny, organized datasets, which makes it difficult for PCM to effectively aggregate large quantities of data, particularly when dealing with diverse data types. The study proposes the use of a high-order PCM method (HOPCM) in large-scale data clustering to effectively address the issue by optimizing the target function within tensor space. In addition, we propose the development of a distributed Hierarchical Optimal Path Centroid Method (HOPCM) technique utilizing the MapReduce framework in order to efficiently manage large and diverse datasets. The BGV encryption system is then combined with HOPCM to produce PPHOPCM, a variant of the HOPCM algorithm that protects privacy. The PPHOPCM algorithm employs approximations for updating the membership matrix and clustering centers, which facilitates the computation of the BGV scheme in a secure manner. On the basis of empirical evidence, it has been determined that PPHOPCM is capable of clustering diverse datasets using cloud computing while maintaining the privacy of private data samples.

According to research published in [4], clustering algorithms have become an important area of study in a number of fields, with a particular emphasis on data mining. Nonetheless, as a result of the vast proliferation of big data applications within the cloud computing environment, these applications are now confronted with a number of obstacles and problems. The large volume of data commonly referred to as "Big Data" poses significant computational challenges for conventional clustering methods. The focus of the study is the efficient management of voluminous data and the production of reliable results at a crucial juncture. Despite ongoing research efforts to develop diverse algorithms to simplify complex clustering methodologies, the management of large datasets continues to be plagued with numerous obstacles. This paper provides a comprehensive examination of the most important clustering algorithms, a comparative analysis of clustering methodologies for managing large-scale data, and a discussion of the fundamental challenges associated with various clustering types. This study's primary objective is to emphasize the main benefits and drawbacks of clustering techniques for effectively managing large-scale data, while also considering a number of other factors.

According to research published in [5], the quantity of data accessible grows daily as a result of recent advances in computer technology. Nonetheless, the abundance of data presents consumers with significant obstacles. Cloud computing services offer a dependable infrastructure for the storage of vast amounts of data. They eradicate a number of prerequisites, such as a

dedicated space and the upkeep of costly computer hardware and software. Massive computer clusters are required for the efficient storage and analysis of vast quantities of data. This article examines the concept, classification, and characteristics of big data, as well as a number of cloud services, including Hortonworks, MapR, Microsoft Azure, Google Cloud, and Amazon Web Services. In addition, a comprehensive comparison of several cloud-based big data frameworks is conducted. Various research obstacles pertaining to distributed database storage, data security, heterogeneity, and data visualization collections are discussed.

The expansion of wireless connectivity, the Internet of Things (IoT), and big data necessitates the development of high-performance data processing tools and algorithms, according to the findings presented in [6]. Since it does not require labeled datasets, data clustering is a valuable analytical technique frequently used to address challenges associated with the Internet of Things (IoT) and big data. It has been demonstrated that metaheuristic algorithms are effective at solving diverse clustering problems. However, the computational costs associated with these algorithms prevent them from effectively managing the vast amounts of data generated by Internet of Things (IoT) devices. This study introduces a novel clustering method based on metaheuristics that utilizes the MapReduce framework to resolve the complexities of large-scale data. The proposed methods utilize the search abilities of a military dog squad to identify optimal centroids, while the MapReduce architecture is used to effectively manage large datasets. By conducting experiments on a set of 17 benchmark functions, the efficiency of the optimization technique proposed in this study is determined. The acquired results are then compared to those of five other cutting-edge algorithms, namely the bat algorithm, particle swarm optimization algorithm, artificial bee colony algorithm, multiverse algorithm, and whale optimization algorithm. This study also demonstrates the use of MapReduce, specifically the MapReduce-based MDBO (MR-MDBO), as a parallel implementation of the proposed technique for clustering the enormous datasets generated by industrial Internet of Things (IoT) applications. In addition, the evaluation of MR-MDBO's efficacy is conducted using three real-world IoT-based datasets provided by industry, along with two benchmark datasets from the UCI repository. MR-MDBO's F-measure and calculation time are compared to those of six other cutting-edge methods. The experimental results demonstrate that the proposed MR-MDBO-based clustering method outperforms the other methods being evaluated in terms of clustering precision and computation delays.

According to the findings of a study [7] conducted by researchers, the aggregation of large-scale data frequently necessitates substantial computational resources. Cloud computing has emerged as an attractive alternative in this regard. In the absence of adequate control measures, however, the costs associated with cloud computing may be substantial. The phenomenon of the long tail has been observed frequently in the context of big data clustering, indicating that a considerable quantity of time is typically devoted to the intermediate and later phases of the grouping procedure. This research aims to minimize the unneeded long tail in the clustering process while maintaining an acceptable level of accuracy at the lowest possible computational cost. A novel strategy is proposed to accomplish cost-effective large-scale data clustering on cloud infrastructure. It is possible to implement a termination condition for the k-means and EM (Expectation-Maximization) algorithms so that they automatically terminate at an early stage when the desired level of accuracy is attained during the training of the regression model using the provided sample data. The results of tests conducted on four well-known data sets indicate that the proposed technique enables k-means and EM algorithms to be implemented in cloud environments

with significant cost-effectiveness. In the context of the conducted case studies, it has been observed that the use of the significantly more efficient k-means algorithm results in the achievement of a 99 percent accuracy level while requiring only 47.71 to 71.14 percent of the computational resources required to achieve 100 percent accuracy. In contrast, the less efficient EM algorithm requires 16.69-32.04 percent more computational resources to achieve the same result. Our approach has the potential to save up to $94,687.49 per use case within the United States land use categorization illustration in order to provide a contextual understanding.

Prior research (8) has examined the implications of geospatial big data, which refers to large datasets containing geographical location information, for identifying urban environments. Due to the necessity of defining appropriate approximation measures and the growing time required for query execution, existing database processing algorithms are unable to provide reliable results in a geographic big data environment in an efficient manner. The functional effects are generated by the clustering technique. However, expanding and accelerating clustering algorithms without jeopardizing their high clustering efficacy remains a significant challenge. The primary contribution of this study is the development of a hierarchical distributed k-medoid clustering algorithm tailored for the processing of geographical queries on large datasets. The proposed model employs the Fuzzy k-Medoids technique to address anomalies in the geographic dataset and deal with data uncertainty, thereby augmenting the efficacy of the k-medoid method and producing more accurate clusters. The employed methodology is complex because it does not depend on the number of appropriate clusters. The proposed model is divided into two distinct phases. Utilizing the parallelism paradigm provided by the Apache Spark framework, local clusters are created using a subset of the entire dataset during the initial phase. Subsequently, in the second stage, the local clusters are merged to produce final clusters that are both compact and trustworthy. By autonomously generating a sufficient number of clusters based on the characteristics of the dataset, the proposed method effectively reduces the quantity of information exchanged during the aggregation process. The results indicate that the proposed model outperforms traditional K-medoids in terms of the precision of derived centers, particularly in applications involving large data.

The use of multiple clusterings has been demonstrated to be advantageous for identifying distinct data patterns that may be obscured when analyzing data from multiple perspectives, as indicated by research published in [9]. Multiple clusterings are beneficial for a variety of applications, including community identification, resource recommendation, and gene expression analysis, among others. Tensor-based Multiple Clustering (TMC) was proposed by the authors as a potential solution to the issue that existing multiple clustering algorithms are primarily designed for low-dimensional data within a single domain and are not well-suited for addressing the challenges posed by Big Data in Cyber-Physical-Social Systems (CPSS). In the event that the size of data continues to grow, however, there will be an exponential increase in the requirements for data storage, processing capabilities, and memory use. The result of this phenomenon will be significant disruptions in spatial dimensions, which will have a significant impact on the overall effectiveness of TMC. This study investigates the application of a TTMC technique and its corresponding parallel computing method. A novel architecture, which utilizes tensor trains (TT) as its foundation, has been created to facilitate a multitude of clustering parallel analytics and services. Afterwards, a suggested strategy to improve the precision and efficacy of TMC is provided. Utilizing a multi-linear attribute combination weight learning technique based on tensor train (TT) representation, a selected weighted tensor train distance, and the TTMC algorithm, this

method is implemented. In addition, TTMC employs TT core parallelism to develop a highly effective distributed parallel computing methodology. Compared to the original TMC method sets, experimental results indicate that the TTMC technique, along with its parallelization, can substantially improve computational efficiency and clustering precision while concurrently reducing memory consumption levels.

According to the findings of reference [10], conventional clustering methods are primarily designed to manage linearly separable data, whereas difficulties arise when attempting to cluster non-linearly separable data within the feature space. This study introduces a novel clustering method, Kernelized Scalable Random Sampling with Iterative Optimization Fuzzy c-Means (KSRSIO-FCM), which is based on a Big Data architecture. Concurrently with the KSRSIO-FCM, the Kernelized Scalable Literal Fuzzy c-Means (KSLFCM) clustering method is introduced. KSLFCM is an essential component of the proposed KSRSIO-FCM algorithm. Using a Radial Basis Function (RBF) kernel, kernelized clustering approaches have been devised to address non-linear separability challenges. This kernel enables the non-linear transformation of the input data space into a higher-dimensional feature space. Our goal is to develop and implement kernelized fuzzy clustering techniques utilizing Apache Spark's in-memory cluster computing technology, which effectively manages massive datasets. For the purpose of demonstrating the efficacy of the proposed KSRSIO-FCM method in comparison to established scalable clustering algorithms such as KSLFCM, SRSIO-FCM, and SLFCM, exhaustive experiments were conducted on diverse sets of massive datasets. According to the experimental results, the KSRSIO-FCM algorithm exhibits significant improvements in time and space complexity, as well as performance metrics such as Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and F-score, when compared to KSLFCM, SRSIO-FCM, and SLFCM. In addition, a comparison of the performance of KSRSIO-FCM and KSLFCM has been conducted. In comparison to conventional scalable fuzzy clustering algorithms, the implementation of KSRSIO-FCM on Apache Spark has greater clustering potential for Big Data samples.

The study conducted by [11] discussed the topic in question. The last several decades have seen a surge in the popularity of distributed computing, mostly due to advancements in clustered computing and big data technologies. The majority of currently used distributed algorithms operate on the premise that the whole of the data is first centralized and then dispersed across several devices. In contemporary times, there is an increasing prevalence of data being distributed across several locations, necessitating the need to perform computations that include the whole of the data while minimizing the associated costs of connectivity. In this study, we propose a novel framework for spectral clustering that facilitates the computation of distant data with less connection requirements, resulting in a substantial acceleration in the calculation process. When comparing the dispersed option to the non-distributed alternative, the decrease in accuracy is negligible. The proposed methodology facilitates the implementation of local parallel computing at the specific location of the data, so leveraging the dispersed nature of the data as an advantageous factor. The maximum acceleration occurs when the data is evenly dispersed across many sites. The conducted experiments on large datasets from UC Irvine, as well as simulated datasets, demonstrate a significant improvement in processing performance by a factor of two when using two distributed sites. Notably, this enhanced speed is achieved without a substantial compromise in accuracy. The architecture we have developed effectively addresses the issue of privacy in data sharing inside distributed computing processes, since it eliminates the need for data to be sent in its original format.

According to the research conducted in reference [12], the provision of realistic mobile network communication services heavily relies on the presence of a wireless communication network (WCN). Random Phase Multiple Access (RPMA) Wireless Communication Network (WCN) is often used in the advancement of wireless networks worldwide due to its advantageous characteristics, including reduced power consumption and high cell density. However, this sort of wireless communication network (WCN) fails to meet the requirements for excellent communication quality in practical applications. The RPMA communication quality prediction model for big data wireless base stations is developed using the lifting regression tree approach and the convolutional neural network (CNN) algorithm. It may be used to ascertain the aspects that distinctly impact the quality of communication. The selection of the convolutional neural network technique was motivated by its capacity to effectively seek and handle nonlinear feature relationships, as well as its relatively low computational complexity. The strategy used is characterized as a black box methodology, hence rendering it unfeasible to get significant coefficients for individual features. Consequently, this limitation poses challenges for conducting further investigations. Therefore, it is crucial to choose a relatively simple modified regression tree method for the computational process. The model for predicting the transmission quality of the RPMA wireless big data base station was developed by combining the convolutional neural network with the lifting regression tree approach. In order to enhance the efficacy of base station deployment, a novel base station planning and deployment model was developed, using the weighted K-centroids algorithm. The significance coefficients of T_B_diff, P_La, and P_Lo in the CNN-DT model exhibit the highest values, significantly surpassing those of other features. Specifically, these coefficients are 0.352, 0.289, and 0.264, respectively. The research conducted in this study has identified that the weighted K-centroids clustering approach exhibits the most superior performance in terms of the distribution of the received signal strength indicator (RSSI) values during downlink reception. The test point proportions for the WK centroids model, K-means model, GMC model, Mean Shift model, and spectral clustering model were 1.95%, 6.25%, 4.25%, 8.22%, and 7.13% correspondingly, for the RSSI bucket range of -140 to -130. The model created in this study, which incorporates a combination of Convolutional Neural Networks (CNN) and an improved regression tree approach, can accurately forecast the communication quality of wireless big data base stations. The main contribution of this study is in its potential to be integrated with the base station deployment and planning model, which is based on the weighted K-centroids algorithm. This integration has the ability to improve the effectiveness and accuracy of base station site selection for the RPMA wireless communication network.

According to the research conducted in reference [13], the precise prediction of power demand plays a crucial role in effectively managing energy systems. Recent efforts have been directed towards enhancing forecasting accuracy via the use of cluster-based methods (CBAs) that include the segmentation of smart meter data into a finite number of clusters, followed by the development of distinct prediction models for each cluster. The total demand is then determined by using the aggregated forecasts based on clustering. Compared to conventional approaches, which often lack compatibility with the integration of data from smart meters, cost-benefit analyses (CBAs) have shown encouraging results. Nevertheless, cost-benefit analyses (CBAs) pose significant computing challenges and are particularly susceptible to dimensionality issues, especially when dealing with large datasets including millions of smart meters. This paper presents a novel strategy known as the reduced model approach (RMA) that effectively integrates fine-resolution, high-dimensional data obtained from millions of smart meters into load prediction. This

integration is achieved via the use of a cutting-edge hierarchical dimension reduction technology. This study utilizes data obtained from a utility company operating in Illinois, United States, which serves a substantial customer base of over 3.7 million individuals. The objective of this research is to demonstrate the effectiveness of our proposed method and evaluate the predictive accuracy of the model used. The implementation of the proposed hierarchical dimension reduction technique enables the expansion of the use of high-resolution data obtained from smart meters, a feat unattainable by other approaches. The results suggest that there are significant improvements in prediction accuracy when compared to current approaches that either lack fine-resolution data or are not suitable for analyzing large-scale smart-meter big data samples.

According to the research conducted in reference [14], distributed computing frameworks serve as the fundamental components of distributed computing systems. They provide a vital mechanism for facilitating the efficient processing of substantial volumes of data on clusters or inside cloud computing environments. The magnitude of big data surpasses the pace at which clusters' capacity for processing it expands. Consequently, distributed computing frameworks that rely on the MapReduce computing paradigm are inadequate for handling big data analysis workloads, which often need the execution of complex analytical algorithms on very large data sets in the terabyte scale. When undertaking these tasks, these frameworks encounter three challenges: limited computational efficiency caused by elevated I/O and communication expenses, memory constraints that hinder the scaling of large datasets, and a dearth of analytical algorithms due to the constraints imposed by the MapReduce programming model on numerous sequential algorithms. In order to address these challenges, it is imperative to develop novel distributed computing frameworks. This paper examines the existing large data management systems, namely those of the MapReduce kind, and evaluates the challenges they provide in the context of big data analysis. In addition, we provide an alternative distributed computing system that diverges from the MapReduce framework, but has promising capabilities in tackling the challenges associated with large-scale data processing.

According to the research conducted in reference [15], federated learning (FL) has gained significant recognition as a technique for facilitating privacy-preserving data exchange within the context of the Internet of Medical Things (IoMT). Furthermore, contemporary scholarly investigations use blockchain technology as a means of safeguarding federated learning (FL). However, current blockchain-based federated learning (BFL) systems encounter difficulties when dealing with sparse data inside a BFL cluster. The optimal approach to constructing a large BFL cluster involves using a multitude of devices. Nevertheless, these devices may be situated in geographically distant and widely separated locations, hence exacerbating the issue of excessive connection latency. The frequent connections in the blockchain consensus lead to a significant latency, which therefore impairs the system efficiency of BFL. This study proposes the subdivision of the main cluster into many smaller clusters, each situated in a unique geographical area and overseen by a BFL (Cluster Management Body). In this particular scenario, we propose the use of CFL, which stands for Cross-Cluster Federated Learning. CFL is a system that operates based on the cross-chain approach. When many BFL clusters are joined via CFL, the system efficiency is enhanced by transmitting a limited quantity of aggregated updates across long distances. The cross-chain consensus approach, which forms the core of the CFL design, guarantees the secure transmission of model modifications between clusters. Thorough experiments are conducted to compare compact fluorescent lamps (CFLs) to traditional incandescent bulbs (BFLs) in order to verify the viability and efficiency of CFLs.

In a study conducted by Work [16], it was highlighted that the processing of matching snapshots of dynamic graphs in dynamic graph analysis applications often requires the creation of several Timing Iterative Graph Processing (TGP) tasks. These tasks are necessary to obtain data at different time periods. The simultaneous execution of TGP processes on the GPU is expected in order to effectively accommodate the high throughput demands of these applications. Despite the recent advancements in GPU-based systems, the processing of dynamic graphs that exceed the GPU's memory capacity is hindered by substantial data access overhead. This is primarily due to the extensive transfer of data between the CPU and GPU, as well as the interference caused by concurrently running tasks. Consequently, the GPU utilization ratio is adversely affected, resulting in suboptimal performance. This research reveals that the TGP tasks exhibit separate snapshots that are used for their individual processing. These snapshots demonstrate significant temporal and spatial similarity, since the majority of the snapshots remain consistent while just a few areas undergo changes throughout time. By substantially reducing the expenses associated with CPU-GPU graph data transfer, it creates optimal circumstances for the efficient simultaneous execution of TGP workloads. In light of this discovery, we have developed EGraph, a novel dynamic graph processing system that utilizes GPU technology. EGraph is designed to seamlessly interact with existing static graph processing systems that operate outside of GPU memory. Its primary objective is to enhance the concurrent execution of TGP (task-based graph processing) operations on dynamic graphs by using the computational power of GPU accelerators. In this research, we propose the implementation of an efficient Loading-Processing-Switching (LPS) execution paradigm inside EGraph. This approach differs from prior methodologies used in this field. The effective execution of TGP activities may be achieved by leveraging the data access commonalities across TGP processes, which reduces the overhead of CPU-GPU data transfer and ensures a greater GPU utilization ratio. Experimental studies indicate that the incorporation of the EGraph method into existing GPU-accelerated systems results in performance improvements ranging from 2.3 to 3.5 times.

In a previous study [17], the author discussed the ongoing development of the Internet of Things (IoTs) and artificial intelligence, which has resulted in the establishment of a novel IoT framework known as the artificial Intelligence of Things (AIoTs). The proliferation of AIoT has led to the accumulation of a substantial volume of unannotated industrial big data. The analysis of large quantities of unlabeled data poses a significant labor and time burden for diagnostic personnel. This research study introduces the deep adaptive fuzzy clustering method (DAFC), which is a novel two-stage unsupervised fault detection system designed to tackle the challenge of unsupervised fault clustering. The unsupervised defect detection approach in the clustering analysis of unlabeled industrial big data involves the integration of stacked sparse autoencoder (SSAE) and adaptive weighted Gath-Geva (AWGG) clustering, known as DAFC. The network may be fine-tuned by SSAE via a two-step process, which involves extracting highly abstract properties from the original data and using different unsupervised approaches. The Gath-Geva clustering algorithm is enhanced by the Adaptive Weighted Genetic Algorithm (AWGG), which has the capability to dynamically get optimal clustering results without requiring a predetermined number of clusters. Based on empirical evidence obtained from two independent datasets, the proposed unsupervised fault detection and clustering (DAFC) method demonstrates its efficacy in accurately identifying fault features from unannotated data. Additionally, this approach exhibits the ability to automatically provide optimal clustering results without prior knowledge of the number of clusters. To the best of our current understanding, this research represents the first endeavor to enhance SSAE without any prior knowledge or data labels. Additionally, it

introduces an unsupervised framework for the purpose of fault detection. The use of Distributed Artificial Intelligence of Things (DAFC) has the potential to serve as a viable solution for implementing collaborative systems in the context of the Internet of Things (IoT) by using large-scale datasets inside industrial settings. Diagnostic specialists use the clustering results generated by DAFC, therefore obviating the need for manual analysis of the unannotated data. This approach offers notable advantages in terms of time efficiency and cost reduction.

Based on the findings in reference [18], it has been observed that a significant number of contemporary graph processing systems use a pull-based computing paradigm to efficiently manage the computationally intensive aspects of graph iteration, hence facilitating high levels of parallelism. Pull models may exhibit a significant number of erroneous vertex and edge operations that do not contribute to graph convergence. This can lead to a drop in performance, since all vertices and edges are scrutinized in every iteration. The findings of this research indicate that a little amount of essential information may be used to effectively eliminate these erroneous processes. Nevertheless, a significant portion of crucial information is often obscured from sight due to the ongoing processing of active vertices. In this study, we propose two distinct filtering methods, namely boundary-cut heuristics and speculative prediction, to effectively uncover hidden critical information in various graph algorithms. These approaches work in a cooperative manner to enhance the discovery process. The integration of three advanced graph processing systems, namely Ligra, Gemini, and Polymer, has been accomplished by combining their respective approaches and developing a hybrid solution. Based on empirical evidence obtained from a diverse collection of graph algorithms applied to both real-world and synthetic graph datasets, it has been shown that none of these strategies can universally guarantee optimal performance. In a variety of use situations, the implementation of boundary-cut, predictive, and hybrid approaches has been seen to enhance performance by 115.1%, 38.1%, and 136.6%, respectively for different scenarios.

Work in [19] discussed that Topological data analysis is a new theoretical trend using topological techniques to mine data. This approach helps determine topological data structures. It focuses on investigating the global shape of data rather than on local information of high-dimensional data. The Mapper algorithm is considered as a sound representative approach in this area. It is used to cluster and identify concise and meaningful global topological data structures that are out of reach for many other clustering methods. In this article, we propose a new method called the Shape Fuzzy C-Means (SFCM) algorithm, which is constructed based on the Fuzzy C-Means algorithm with particular features of the Mapper algorithm. The SFCM algorithm can not only exhibit the same clustering ability as the Fuzzy C-Means but also reveal some relationships through visualizing the global shape of data supplied by the Mapper. We present a formal proof and include experiments to confirm our claims. The performance of the enhanced algorithm is demonstrated through a comparative analysis involving the original algorithm, Mapper, and the other fuzzy set based improved algorithm, F-Mapper, for synthetic and real-world data. The comparison is conducted with respect to output visualization in the topological sense and clustering stability levels.

Work in [20] discussed that The awareness of edge computing is attaining eminence and is largely acknowledged with the rise of the Internet of Things (IoT). Edge-enabled solutions offer efficient computing and control at the network edge to resolve the scalability and latency-related concerns. Though, it comes to be challenging for edge computing to tackle diverse applications of IoT as they produce massive heterogeneous data. The IoT-enabled frameworks for

Big Data analytics face numerous challenges in their existing structural design, for instance, the high volume of data storage and processing, data heterogeneity, and processing time among others. Moreover, the existing proposals lack effective parallel data loading and robust mechanisms for handling communication overhead. To address these challenges, we propose an optimized IoT-enabled big data analytics architecture for edge–cloud computing using machine learning. In the proposed scheme, an edge intelligence module is introduced to process and store the big data efficiently at the edges of the network with the integration of cloud technology. The proposed scheme is composed of two layers: 1) IoT–edge and 2) cloud processing. The data injection and storage is carried out with an optimized MapReduce parallel algorithm. An optimized yet another resource negotiator (YARN) is used for efficiently managing the cluster. The proposed data design is experimentally simulated with an authentic data set using Apache Spark. The comparative analysis is decorated with the existing proposals and traditional mechanisms. The results justify the efficiency of our proposed work process.

Work in [21] discussed that Many iterative graph processing systems have recently been developed to analyze graphs. Although they are effective from different aspects, there is an important issue that has not been addressed yet. A real-world graph follows the power-law property, in which a small number of vertices have high degrees (i.e., are connected to most other vertices in the graph). These vertices are called hot-vertices and usually require more iterations to converge. In the existing solutions, these hot-vertices may be allocated to many or even all graph partitions along with other vertices that are easy to converge. As the result, the partitions with hot-vertices have to be loaded repeatedly (and consequently the system suffers from high data access cost), although perhaps only a few vertices in these partitions are active. To cope with this issue, we develop an efficient open source graph partition manager, called GGraph, which can be integrated into the existing graph processing systems to efficiently support iterative graph processing, by taking into account the power-law property of the graph structure. It uses a novel graph repartitioning scheme with low overhead to dynamically partition the hot-vertices together, so as to avoid loading the inactive vertices in the same partition as the repeatedly processed hot-vertices. By such means, it not only enables less data access cost, but also enables the privileged processing of the hot-vertices. In order to further increase the convergence speed, a scheduling algorithm is further proposed in this work to prioritize the processing of the hot-vertices with low overhead. To demonstrate the efficiency of GGraph, we plug it into four state-of-the-art graph processing systems, i.e., Gemini, GraphChi, Chaos, and GridGraph, and experimental results show that GGraph improves their performance by up to 3.2 times, 3.8 times, 3.9 times, 3.5 times, respectively for different scenarios.

Work in [22] discussed that Large-scale data clustering is an essential key for big data problem. However, no current existing approach is “optimal” for big data due to high complexity, which remains it a great challenge. In this article, a simple but fast approximate DBSCAN, namely, KNN-BLOCK DBSCAN, is proposed based on two findings: 1) the problem of identifying whether a point is a core point or not is, in fact, a kNN problem and 2) a point has a similar density distribution to its neighbors, and neighbor points are highly possible to be the same type (core point, border point, or noise). KNN-BLOCK DBSCAN uses a fast approximate kNN algorithm, namely, FLANN, to detect core-blocks (CBs), noncore-blocks, and noise-blocks within which all points have the same type, then a fast algorithm for merging CBs and assigning noncore points to proper clusters is also invented to speedup the clustering process. The experimental results show that KNN-BLOCK DBSCAN is an effective approximate DBSCAN algorithm with high accuracy, and outperforms other

current variants of DBSCAN, including ρ-approximate DBSCAN and AnyDBC process.

Work in [23] discussed that Life pattern clustering is essential for abstracting the groups' characteristics of daily life patterns and activity regularity. Based on millions of GPS records, this research proposes a framework on the life pattern clustering which can efficiently identify the groups that have similar life patterns. The proposed method can retain original features of individual life pattern data without aggregation. Metagraph-based data structure is proposed for presenting the diverse life pattern. Spatial-temporal similarity includes significant places semantics, time-sequential properties and frequency are integrated into this data structure, which captures the uncertainty of an individual and the diversities between individuals. Non-negative-factorization-based method is utilized for reducing the dimension. The results show that our proposed method can effectively identify the groups that have similar life pattern in long term and takes advantage in computation efficiency and representational capacity compared with the traditional methods. We reveal the representative life pattern groups and analyze the group characteristics of human life patterns during different periods and different regions. We believe our work helps in future infrastructure planning, services improvement and policy making related to urban and transportation, thus promoting a humanized and sustainable city sets.

Work in [24] discussed that Matrix decomposition is one of the fundamental tools to discover knowledge from big data generated by modern applications. However, it is still inefficient or infeasible to process very big data using such a method in a single machine. Moreover, big data are often distributedly collected and stored on different machines. Thus, such data generally bear strong heterogeneous noise. It is essential and useful to develop distributed matrix decomposition for big data analytics. Such a method should scale up well, model the heterogeneous noise, and address the communication issue in a distributed system. To this end, we propose a distributed Bayesian matrix decomposition model (DBMD) for big data mining and clustering. Specifically, we adopt three strategies to implement the distributed computing including 1) the accelerated gradient descent, 2) the alternating direction method of multipliers (ADMM), and 3) the statistical inference. We investigate the theoretical convergence behaviors of these algorithms. To address the heterogeneity of the noise, we propose an optimal plug-in weighted average that reduces the variance of the estimation. Synthetic experiments validate our theoretical results, and real-world experiments show that our algorithms scale up well to big data and achieves superior or competing performance compared to two typical distributed methods including Scalable-NMF and scalable k-means++ process.

Work in [25] discussed that Change is one of the biggest challenges in dynamic stream mining. From a data-mining perspective, adapting and tracking change is desirable in order to understand how and why change has occurred. Clustering, a form of unsupervised learning, can be used to identify the underlying patterns in a stream. Density-based clustering identifies clusters as areas of high density separated by areas of low density. This paper proposes a Multi-Density Stream Clustering (MDSC) algorithm to address these two problems; the multi-density problem and the problem of discovering and tracking changes in a dynamic stream. MDSC consists of two on-line components; discovered, labelled clusters and an outlier buffer. Incoming points are assigned to a live cluster or passed to the outlier buffer. New clusters are discovered in the buffer using an ant-inspired swarm intelligence approach. The newly discovered cluster is uniquely labelled and added to the set of live clusters. Processed data is subject to an ageing function and will disappear when it is no longer relevant. MDSC is shown to perform favourably to state-of-the-art peer

stream-clustering algorithms on a range of real and synthetic data-streams. Experimental results suggest that MDSC can discover qualitatively useful patterns while being scalable and robust to noise levels. Thus, a wide variety of models are proposed for improving clustering in big data applications. An empirical comparison of these models is discussed in the next section of this text.

III.Result analysis & comparison

Based on the in-depth analysis of different models used for analysis of big data clustering, it can be observed that these models are highly variant in terms of their functional characteristics. To further contemplate this analysis, an empirical comparison of these models is tabulated in table 1, where these models are compared in terms of Accuracy (A), Computational Complexity (CC), Scalability (S), and Applicability (App) to multiple applications. While accuracy is compared in terms of absolute values, other metrics are converted into fuzzy ranges of Low (1), Medium (2), High (3), and Very High (VH) levels.

METHOD	A	CC	S	APP
SFCM [1]	99.67	4	1	1
IOT [2]	90.75	1	3	4
GGRAPH [3]	95.93	2	2	4
KNN-BLOCK DBSCAN [4]	90.96	4	2	1
MDSC [5]	98.4	3	4	4
DAFC [6]	94.97	2	2	1
CFL [7]	96.06	3	3	2
RMA [8]	98.16	2	1	1
EGRAPH [9]	97.4	3	1	3
TGP [10]	99.8	2	4	4
CBAS [11]	96.46	2	3	2
RPMA [12]	91.77	2	1	1
FL [13]	99.41	4	3	3
ADMM [14]	97.52	4	1	2
DBMD [15]	95.85	3	3	3
MAPPER [16]	94.8	3	3	1
WCNA [17]	93.39	3	1	2
AIOTS [18]	94.2	3	1	1
MCR [19]	90.16	2	4	3
DFSM [20]	95.34	2	1	4
MCDM [21]	93.84	1	1	2
SSDA [22]	94.12	2	3	4
IAI [23]	92.31	3	4	1
TPSA [24]	95.77	1	4	1
WKMF [25]	93.28	1	3	3

Table 1. Comparative Analysis of Different Models

In terms of precision, the examined methodologies exhibit varying degrees of effectiveness. In this study, the leading candidates are TGP [10], FL [13], and SFCM [1], with respective accuracy rates of 99.8%, 99.41%, and 99.67%. The MDSC [5], RMA [8], ADMM [14], EGraph [9], and CFL [7] have moderately accurate accuracy values ranging from 98.4% to 96.06%. The results of GGraph [3], DBMD [15], SSDA [22], DAFC [6], and IAI [23] indicate a significant decrease in accuracy, ranging from 95.93% to 92.3%. Several investigations, including WKMF [25], WCNA [17], MCDM [21], MAPPER [16], and AIOTS [18], have demonstrated a decline in precision spanning from 94.2% to 93.28%.

Considering the computational characteristics of the models, a spectrum of computational complexity is evident. Low computational complexity algorithms, such as RMA [8], AIOTS [18], RPMA [12], and WKMF [25], have been devised to process data efficiently while minimizing computational burden. Moderately complex models, including DAFC (6), CFL (7), GGraph (3), KNN-BLOCK DBSCAN (4), MCR (19), MDSC (5), and TPSA (24), balance processing efficiency and complexity. Due to their advanced algorithms and methodologies, SFCM [1], MDSC [5], EGraph [9], CBAS [11], DAFC [6], IOT [2], and TGP [10] incur significant computational costs.

The models contain a variety of magnitude levels in terms of their scalability. TGP [10], SFCM [1], and

MDSC [5] are the systems that demonstrate exceptional scalability in managing growing data volumes. GGraph [3], CFL [7], DAFC [6], EGraph [9], ADMM [14], DBMD [15], and SSDA [22] can accommodate expanding datasets to a limited extent due to their limited scalability. On the other hand, the limited scalability demonstrated by RPMA [12], WCNA [17], AIOTS [18], MAPPER [16], IAI [23], MCDM [21], WKMF [25], MCR [19], KNN-BLOCK DBSCAN [4], and TPSA [24] indicates limitations in managing larger datasets effectively.

The models are applicable in multiple domains and accommodate varying degrees of competence. TGP (reference [10]), MDSC (reference [5]), SFCM (reference [1]), CFL (reference [7]), DBMD (reference [15]), and IAI (reference [23]) are versatile options in numerous contexts due to their pervasive application (references [10], [5], [7], and [15]). On the other hand, it should be noted that GGraph [3], RPMA [12], WCNA [17], AIOTS [18], TPSA [24], MDSC [5], ADMM [14], DAFC [6], EGraph [9], KNN-BLOCK DBSCAN [4], MCDM [21], WKMF [25], MCR [19], MAPPER [16], SSDA [22], IOT [2], and CBAS [11] exhibit distinct applicability that is specifically tailored to various specialized domains, such as graph processing, wireless communication, IoT analytics, and other relevant contexts.

IV.Conclusion and future scope

I Through this exhaustive analysis, we have gained a substantial understanding of the efficacy, complexity, scalability, and applicability of 25 distinct data mining and analytics approaches. The analysis presented in this study emphasizes the benefits and specialized knowledge associated with each approach, providing researchers and practitioners with a guide for making informed decisions based on their particular requirements.

The high accuracy rates demonstrated by TGP, FL, and SFCM have established them as formidable competitors in precision-driven occupations, emphasizing the significance of accuracy as a factor that must be considered. These models have the ability to recognize intricate patterns within datasets, making them valuable tools for applications requiring high precision.

Variation in computational complexity was observed between models. Models like RMA, AIOTS, RPMA, and WKMF have minimal computational complexity and effectively balance performance and efficiency. In contrast, models such as SFCM, MDSC, EGraph, CBAS, DAFC, IOT, and TGP are capable of conducting more intricate data analysis duties due to their increased complexity.

The scalability characteristic demonstrates the capacity of models to adapt to expanding datasets. TGP, SFCM, and MDSC demonstrate extraordinary performance in managing large volumes of data, making them ideally suited for scenarios involving voluminous quantities of data. Nonetheless, some models exhibit insufficient scalability, highlighting the importance of selecting the appropriate instrument based on the size of the dataset samples.

These methods have a wide variety of applications, ranging from general applicability to specialized domains. Several models, such as TGP, MDSC, SFCM, CFL, DBMD, and IAI, are applicable to a wide range of endeavors. In contrast, other models, such as graph processing, wireless communication, IoT analytics, and others, are purpose-built for specific duties. The diverse range of applicability emphasizes the need to painstakingly align the strategy with the application's particular scenarios.

Future Scope:

This study provides a conceptual framework that could serve as a foundation for future research and development in the field of data mining and analytics. Researchers have the opportunity to conduct

additional research on these findings in order to address particular issues and explore novel avenues. The following are prospective research areas for the future:

1. Hybrid Models: Examine the viability of combining the merits of multiple approaches to produce hybrid models that combine the accuracy of some techniques with the computational efficiency of others, resulting in improved performance levels.

Enhancements to Scalability: Develop methodologies to enhance the scalability of models with limited scalability, allowing them to process larger datasets and expanding their potential applications.

Resource Optimization: It is crucial to focus on improving the efficacy and availability of computing resources required by complex models in order to expand their applicability to a broader range of contexts.

4. Real-time Applications: Analyze the potential use of these methodologies in real-time data streaming contexts, allowing for swift comprehension and decision-making in volatile situations.

5. Interdisciplinary Applications: Utilize these methodologies in interdisciplinary disciplines such as healthcare, economics, and environmental monitoring to extract valuable information regarding hidden patterns and tendencies.

Perform exhaustive comparisons of a large number of models in order to assess their efficacy in specific scenarios. This will help practitioners choose the optimal model for their data analysis duties.

Explainable artificial intelligence (AI) refers to the improvement of AI model transparency and interpretability. This enables individuals to comprehend the rationale underlying the generated insights, thereby nurturing trust and confidence in the outcomes.

8. Development of Automated Model Selection: Design and implement software solutions that use automated algorithms to select the optimal model for a given dataset based on its inherent characteristics. This would facilitate analysts' decision-making and streamline their work processes.

Therefore, this research functions as a valuable resource for both academics and industry professionals, facilitating in the determination of optimal data mining and analytics methods by taking into account factors such as precision, computational complexity, scalability, and practical applicability levels. The future prospects indicated provide a strategic plan for advancing innovations in the field, thereby facilitating the growth and development of data-driven solutions across numerous industries & scenarios.

V. References

[1] M. S. Mahmud, J. Z. Huang, R. Ruby, A. Ngueilbaye and K. Wu, "Approximate Clustering Ensemble Method for Big Data," in IEEE Transactions on Big Data, vol. 9, no. 4, pp. 1142-1155, 1 Aug. 2023, doi: 10.1109/TBDATA.2023.3255003.

[2] Z. Hu and D. Li, "Improved heuristic job scheduling method to enhance throughput for big data analytics," in Tsinghua Science and Technology, vol. 27, no. 2, pp. 344-357, April 2022, doi: 10.26599/TST.2020.9010047.

[3] Q. Zhang, L. T. Yang, Z. Chen and P. Li, "PPHOPCM: Privacy-Preserving High-Order Possibilistic c-Means Algorithm for Big Data Clustering with Cloud Computing," in IEEE Transactions on Big Data, vol. 8, no. 1, pp. 25-34, 1 Feb. 2022, doi: 10.1109/TBDATA.2017.2701816.

[4] M. A. Mahdi, K. M. Hosny and I. Elhenawy, "Scalable Clustering Algorithms for Big Data: A Review," in IEEE Access, vol. 9, pp. 80015-80027, 2021, doi: 10.1109/ACCESS.2021.3084057.

[5] A. K. Sandhu, "Big data with cloud computing: Discussions and challenges," in Big Data Mining

and Analytics, vol. 5, no. 1, pp. 32-40, March 2022, doi: 10.26599/BDMA.2021.9020016.

[6] A. K. Tripathi, K. Sharma, M. Bala, A. Kumar, V. G. Menon and A. K. Bashir, "A Parallel Military-Dog-Based Algorithm for Clustering Big Data in Cognitive Industrial Internet of Things," in IEEE Transactions on Industrial Informatics, vol. 17, no. 3, pp. 2134-2142, March 2021, doi: 10.1109/TII.2020.2995680.

[7] D. Li, S. Wang, N. Gao, Q. He and Y. Yang, "Cutting the Unnecessary Long Tail: Cost-Effective Big Data Clustering in the Cloud," in IEEE Transactions on Cloud Computing, vol. 10, no. 1, pp. 292-303, 1 Jan.-March 2022, doi: 10.1109/TCC.2019.2947678.

[8] M. M. Madbouly, S. M. Darwish, N. A. Bagi and M. A. Osman, "Clustering Big Data Based on Distributed Fuzzy K-Medoids: An Application to Geospatial Informatics," in IEEE Access, vol. 10, pp. 20926-20936, 2022, doi: 10.1109/ACCESS.2022.3149548.

[9] Y. Zhao et al., "Tensor Train-Based Multiple Clusterings for Big Data in Cyber-Physical-Social Systems and Its Efficient Implementations," in IEEE Transactions on Network Science and Engineering, vol. 9, no. 6, pp. 3896-3908, 1 Nov.-Dec. 2022, doi: 10.1109/TNSE.2021.3119324.

[10] P. Jha, A. Tiwari, N. Bharill, M. Ratnaparkhe, M. Mounika and N. Nagendra, "A Novel Scalable Kernelized Fuzzy Clustering Algorithms Based on In-Memory Computation for Handling Big Data," in IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 5, no. 6, pp. 908-919, Dec. 2021, doi: 10.1109/TETCI.2020.3016302.

[11] D. Yan, Y. Wang, J. Wang, G. Wu and H. Wang, "Fast Communication-Efficient Spectral Clustering over Distributed Data," in IEEE Transactions on Big Data, vol. 7, no. 1, pp. 158-168, 1 March 2021, doi: 10.1109/TBDATA.2019.2907985.

[12] X. He, T. Yu, Y. Shen and S. Wang, "Traffic Processing Model of Big Data Base Station Based on Hybrid Improved CNN Algorithm and K-Centroids Clustering Algorithm," in IEEE Access, vol. 11, pp. 63057-63068, 2023, doi: 10.1109/ACCESS.2023.3286860.

[13] N. Alemazkoor, M. Tootkaboni, R. Nateghi and A. Louhghalam, "Smart-Meter Big Data for Load Forecasting: An Alternative Approach to Clustering," in IEEE Access, vol. 10, pp. 8377-8387, 2022, doi: 10.1109/ACCESS.2022.3142680.

[14] X. Sun, Y. He, D. Wu and J. Z. Huang, "Survey of Distributed Computing Frameworks for Supporting Big Data Analysis," in Big Data Mining and Analytics, vol. 6, no. 2, pp. 154-169, June 2023, doi: 10.26599/BDMA.2022.9020014.

[15] H. Jin, X. Dai, J. Xiao, B. Li, H. Li and Y. Zhang, "Cross-Cluster Federated Learning and Blockchain for Internet of Medical Things," in IEEE Internet of Things Journal, vol. 8, no. 21, pp. 15776-15784, 1 Nov.1, 2021, doi: 10.1109/JIOT.2021.3081578.

[16] Y. Zhang et al., "EGraph: Efficient Concurrent GPU-Based Dynamic Graph Processing," in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 6, pp. 5823-5836, 1 June 2023, doi: 10.1109/TKDE.2022.3171588.

[17] X. Hu, Y. Li, L. Jia and M. Qiu, "A Novel Two-Stage Unsupervised Fault Recognition Framework Combining Feature Extraction and Fuzzy Clustering for Collaborative AIoT," in IEEE Transactions on Industrial Informatics, vol. 18, no. 2, pp. 1291-1300, Feb. 2022, doi: 10.1109/TII.2021.3076077.

[18] L. Zheng et al., "Efficient Graph Processing with Invalid Update Filtration," in IEEE Transactions on Big Data, vol. 7, no. 3, pp. 590-602, 1 July 2021, doi: 10.1109/TBDATA.2019.2921358.

[19] Q. -T. Bui et al., "SFCM: A Fuzzy Clustering Algorithm of Extracting the Shape Information of Data," in IEEE Transactions on Fuzzy Systems, vol. 29, no. 1, pp. 75-89, Jan. 2021, doi: 10.1109/TFUZZ.2020.3014662.

[20] M. Babar, M. A. Jan, X. He, M. U. Tariq, S. Mastorakis and R. Alturki, "An Optimized IoT-Enabled Big Data Analytics Architecture for Edge–Cloud Computing," in IEEE Internet of Things Journal, vol. 10, no. 5, pp. 3995-4005, 1 March1, 2023, doi: 10.1109/JIOT.2022.3157552.

[21] B. Si et al., "GGraph: An Efficient Structure-Aware Approach for Iterative Graph Processing," in IEEE Transactions on Big Data, vol. 8, no. 5, pp. 1182-1194, 1 Oct. 2022, doi: 10.1109/TBDATA.2020.3019641.

[22] Y. Chen et al., "KNN-BLOCK DBSCAN: Fast Clustering for Large-Scale Data," in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 51, no. 6, pp. 3939-3953, June 2021, doi: 10.1109/TSMC.2019.2956527.

[23] W. Li et al., "Metagraph-Based Life Pattern Clustering With Big Human Mobility Data," in IEEE Transactions on Big Data, vol. 9, no. 1, pp. 227-240, 1 Feb. 2023, doi: 10.1109/TBDATA.2022.3155752.

[24] C. Zhang, Y. Yang, W. Zhou and S. Zhang, "Distributed Bayesian Matrix Decomposition for Big Data Mining and Clustering," in IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 8, pp. 3701-3713, 1 Aug. 2022, doi: 10.1109/TKDE.2020.3029582.

[25] C. Fahy and S. Yang, "Finding and Tracking Multi-Density Clusters in Online Dynamic Data Streams," in IEEE Transactions on Big Data, vol. 8, no. 1, pp. 178-192, 1 Feb. 2022, doi: 10.1109/TBDATA.2019.2922969.

[26] Shivadekar, S., Kataria, B., Limkar, S. et al. Design of an efficient multimodal engine for preemption and post-treatment recommendations for skin diseases via a deep learning-based hybrid bioinspired process. Soft Comput (2023).

[27] Shivadekar, Samit, et al. "Deep Learning Based Image Classification of Lungs Radiography for Detecting COVID-19 using a Deep CNN and ResNet 50." International Journal of Intelligent Systems and Applications in Engineering 11.1s (2023): 241-250.

[28] P. Nguyen, S. Shivadekar, S. S. Laya Chukkapalli and M. Halem, "Satellite Data Fusion of Multiple Observed XCO2 using Compressive Sensing and Deep Learning," IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 2020, pp. 2073-2076, doi: 10.1109/IGARSS39084.2020.9323861.

[29] Banait, Satish S., et al. "Reinforcement mSVM: An Efficient Clustering and Classification Approach using reinforcement and supervised Techniques." International Journal of Intelligent Systems and Applications in Engineering 10.1s (2022): 78-89.

[30] Shewale, Yogita, Shailesh Kumar, and Satish Banait. "Machine Learning Based Intrusion Detection in IoT Network Using MLP and LSTM." International Journal of Intelligent Systems and Applications in Engineering 11.7s (2023): 210-223.

[31] Vanjari, Hrishikesh B., Sheetal U. Bhandari, and Mahesh T. Kolte. "Enhancement of Speech for Hearing Aid Applications Integrating Adaptive Compressive Sensing with Noise Estimation Based Adaptive Gain." International Journal of Intelligent Systems and Applications in Engineering 11.7s (2023): 138-157.

[32] Vanjari, Hrishikesh B., and Mahesh T. Kolte. "Comparative Analysis of Speech Enhancement Techniques in Perceptive of Hearing Aid Design." Proceedings of the Third International Conference on Information Management and Machine Intelligence: ICIMMI 2021. Singapore: Springer Nature Singapore, 2022.

Cite this article as :

Trishali Dhote, Prof. Pragati Patil, "An in-Depth Review of Big Data Analytic Models for Clustering Operations ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 5, pp.30-47, September-October-2023.