JETIR.ORG
ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

“Distributed Data Mining Methods & Strategies: An In-Depth Analysis”

[1] Amit Ahalawat, [2] Mr . Arun Kumar Jhapate
[1]M.Tech Scholar, 2 Assistant Professor.

[1]Department of Computer Science and Data Science
[1] Sagar Institute of Research and Technology,
Bhopal (M.P.), India

ABSTRACT

Distributed systems, which may be used to do calculations, are being created in response to the rapid development of resource sharing. Data mining, which has a wide variety of real-world applications, offers substantial tools for extracting meaningful and useable information from vast datasets. Traditional data mining approaches, on the other hand, assume that the data is collected centrally, kept in memory, and remains static. Managing and analyzing large volumes of data with limited resources is tough. Large amounts of data, for example, are quickly created and saved in several places. It gets more expensive to consolidate them in one spot. Furthermore, conventional data mining techniques often have a number of challenges and limits, including memory constraints, restricted computing power, and inadequate hard drive capacity, among others. To address the following challenges, distributed data mining has emerged as a viable solution in various applications. According to several authors, this study examines cutting-edge distributed data mining methods such as distributed common item-set mining, distributed frequent sequence mining, technical challenges with distributed systems, distributed clustering, and privacy-protected distributed data mining. Furthermore, each work is reviewed and compared to its peers.

Keywords: DISTRIBUTED SYSTEMS, DATA MINING, PARALLEL PLATFORM, DISTRIBUTED CLUSTERING.

1. INTRODUCTION

Knowledge Discovery in Databases (KDD) is a sophisticated tool that utilizes the advancements in information technology and data collecting to extract valuable and relevant information from a dataset [1]. KDD, or Knowledge Discovery in Databases, has a wide range of practical uses and has led to the development of several Data Mining tasks, including Association Rule Mining (ARM), Sequential Pattern Mining (SPM), Clustering, Classification, and Outlier Detection, among others [2]. The information that is found may be divided into multiple categories depending on the needs in various domains and applications. These categories include frequent item sets and association rules, sequential patterns, sequential rules, graphs, high utility patterns, weight-based patterns, and other interesting patterns [3]. Frequent item-set mining (FIM) or association rule mining (ARM) has been intensively researched because to its significance in several real-world applications [4]. As a result of the fast expansion of resource sharing, distributed systems have been created to efficiently use calculations [5]. Data mining (DM) is a very effective method for extracting valuable and practical information from vast quantities of data. It has a diverse variety of practical applications in many fields. Conversely, conventional DM algorithms use the assumption that the data is gathered in a centralized manner, stored in memory, and remains unchanged [6]. Managing large-scale data and processing them with limited resources is a formidable task. As an example, substantial quantities of data are rapidly generated and stored at several sites [7]. The cost of centralizing them at a single location is rising. In addition, conventional DM algorithms often encounter issues and obstacles, such as memory constraints, limited computing power, and insufficient storage capacity, among others [8]. In order to address the aforementioned issues, the use of DM in distributed computing environments, also known as distributed data mining (DDM), has been gaining prominence as a viable option in several applications [9]. Computation and data distribution are powerful tools that allow engineers and researchers to address a broad range of challenges. These tools may be used and utilized in many distributed applications [10]. Distributed systems refer to linked processing units that are structured via networks to meet the requirements of large-scale or high-performance computing. This field has gained considerable interest in recent decades [11,12].Peer-to-peer (P2P) systems, grids, ad-hoc networks, cloud computing systems, and social network systems have all garnered significant interest. Currently, distributed systems are used for many reasons such as internet services, file storage, and scientific computation [16]. Recently, there has been a rise in the use of centralized data mining methods to analyze vast quantities of business or scientific data stored in databases [17]. The primary concern in data mining is the prompt and precise determination of the correlation between data sets [18]. The rise of vast and enormous data requires the use of a single computer to carry out the calculation process [19]. Nevertheless, the gradual surge in data volume compels academics to devise more intricate ways or tactics to tackle this difficulty [20]. In recent years, there has been a surge in the use of distributed computing and parallel processing, particularly in the field of data analysis and knowledge extraction [21]. The advent of distributed computing some years ago may have addressed the current situation, whereby the volume of mining data has expanded from the size of Megabytes to Gigabytes to currently include Terabytes to Petabytes [22]. On a daily basis, social media and network services produce an enormous volume of data that reaches the Petabyte scale [23]. Due to the existence of large datasets and the need for fast information retrieval, the use of parallel or distributed computing is very pertinent in the present day [24]. Clusters may now easily connect commodity hardware to perform sophisticated functions in a distributed system [25]. Utilizing distributed computing with data mining may significantly improve the efficiency of data mining algorithms, particularly when dealing with dispersed and extensive datasets [26]. The rise of distributed data mining has lately gained great importance. The subject matter of this field involves the manipulation of data in a distributed system, with a particular focus on various aspects related to computing, storage, data sharing, and human-computer interaction [27]. Concurrently, data mining has been extensively investigated [26].

Data mining methods may be used by organizations, enterprises, industries, and research centers to reveal various forms of confidential but useful and important patterns and information [28]. As mentioned earlier, data mining methods may be used to analyze the distribution of gathered data [29]. A notable scenario in data mining occurs when many entities share datasets, with each entity possessing a portion of the data. In the past, established approaches operated under the assumption that the data was concentrated and stored in memory [30]. This assumption is no longer applicable in dispersed networks [31]. Regrettably, the application of traditional mining methods directly to distributed databases proves to be inefficient because of the substantial communication cost [32]. Therefore, the integration of high-performance data mining in distributed computing systems has significantly improved the scalability of these systems [33].

Traditional data mining tools are not suitable for a centralized approach due to many reasons. These include the large amount of data, the difficulty to centralize data from different places, bandwidth restrictions, energy constraints, and privacy concerns [21,34]. Recently, distributed data mining has evolved as a prominent study subject to address these difficulties. In the field of distributed data mining research, there are two often used theories about the distribution of data over different sites: heterogeneously (also known as vertical partitioning) or homogeneously (also known as horizontal partitioning) [35].

Distributed data mining addresses several challenges in analyzing distributed data and offers algorithmic methods for doing different mining operations and data analysis in a distributed way that takes into account resource limitations [36]. Researchers use many methodologies to manage distributed systems such as cloud, grid computing, Hadoop [37, 38], and others. They distribute the mining computation over multiple nodes to enhance the efficiency and scalability of data mining [39]. Prior studies have shown that distributed data mining is a viable approach for end-users, governments, or enterprises to analyze data and discover different forms of important information [40]. It presents novel opportunities while also presenting some challenges for data mining. How can we effectively condense and categorize various research studies on data mining in distributed systems? The objective of this study is to examine the existing studies on data mining of distributed systems [41]. The primary contributions of this study are doing a comprehensive evaluation of current data mining studies conducted on distributed systems [42]. This work focuses on the analysis of distributed system techniques for data mining in several domains, such as distributed clustering and the protection of privacy in distributed data mining [43].

Distributed systems provide a significant gain in productivity by using extensive parallelism [44]. Regrettably, augmenting the quantity of computers accessible to clients would concurrently amplify the need for troubleshooting in instances of malfunction [45]. Data mining allows for the identification and extraction of patterns from large amounts of data, making it a valuable method for troubleshooting distributed systems [46].
Data mining offers methods for identifying correlations, trends, and insights in large datasets, which may be used to make informed decisions regarding future activities. This approach has been used by other disciplines to achieve efficient data processing [47]. When dealing with large and geographically dispersed datasets, there are various challenges that need to be overcome [48].

The proliferation of computers and advancements in database technology has resulted in a significant influx of data [49]. The exponential growth of data stored in databases has created a need to seek knowledge and experience in efficient data mining methods [50]. However, there is a growing need for scalable data mining tools because to the development of Network Distribution Computing, which includes restricted intranet, internet, and wireless networks [51]. Distribution Data Mining (DDM) aims to discover and integrate information from several data sources that are geographically dispersed across various places. Nevertheless, the use of data mining methods for such systems is accompanied by many issues [52]. A distributed computing system is inherently more intricate than a central or host-based system. Dealing with heterogeneous systems, several databases, and potentially different strategies may be required [53]. The communications protocol between nodes should possess scalability and efficacy, while also considering the selective use of information gathered from different nodes [54].

DDM encounters a significant challenge in creating mining algorithms that minimize unnecessary data communication [55]. This skill is required for the purpose of enhancing efficiency and ensuring precise results while maintaining privacy. In addition, the extraction of scattered data necessitates the implementation of appropriate protocols, languages, and network services to effectively handle the required information and mapping [56].

The research is structured in the following manner: The section on technical problems in Distributed Systems introduces the terminology and key elements of distributed systems. The section on Data Mining Techniques in Distributed Environment focuses on the latest advancements and provides an in-depth analysis.

2. OVERVIEW OF DISTRIBUTED SYSTEMS AND DATA MINING

Distributed computing is a branch of computer science that focuses on the study of distributed systems [54]. A distributed system is a networked system where the components are spread over several computers. These components interact and coordinate their activities by exchanging messages with each other from any part of the system [57]. The components collaborate with each other to accomplish a shared objective. The three key characteristics of distributed systems are the simultaneous operation of components, the absence of a universal time reference, and the individual failure of components [58]. Instances of distributed systems range from service-oriented architecture (SOA) based systems to massively multiplayer online games to peer-to-peer applications [59]. A distributed program refers to a computer program that operates inside a distributed system. Distributed programming, on the other hand, is the act of developing such programs [59]. Various implementations of the message transmission mechanism exist, such as pure Hypertext Transfer Protocol (HTTP), RPC-like connectors, and message queues [60]. Distributed computing is the use of distributed systems to address computational challenges. In the field of distributed computing, a given issue is partitioned into several jobs, each of which is resolved by one or more computers. These machines establish communication with one another by message forwarding [39].

The term "distributed" in phrases like "distributed system", "distributed programming", and "distributed algorithm" first denoted computer networks in which individual computers were physically dispersed throughout a certain geographic region [61]. Currently, these phrases are used in a broader context, including autonomous processes that operate on the same physical computer and communicate with each other via message passing. Although a distributed system does not have a universally accepted definition, it is typically characterized by the following properties:

- There are several independent computing entities, such as computers or nodes, each with its own private local memory [62].
- The entities engage in communication via the exchange of messages.

A distributed system is a network of autonomous processors that work together to achieve a shared purpose, such as solving a complex computing issue. From the user's perspective, these processors are seen as a unified entity. Alternatively, each computer may have its own dedicated user with unique requirements, and the objective of the distributed system is to synchronize the use of shared resources or provide communication services to the users [63].

Additional common characteristics of distributed systems include:

- The system's structure (network topology, network latency, number of computers) is not predetermined, and the system may include various types of computers and network links. Additionally, the system may undergo changes while a distributed program is running [64].
- The system must be able to handle failures in individual computers.
- Each computer has a restricted and imperfect perspective of the system. Each computer is limited to have knowledge of just a single portion of the input.

Distributed computing employs a range of diverse hardware and software architectures. At a more basic level, it is essential to link many CPUs via a network, whether that network is integrated into a circuit board or consists of separate devices and connections [65]. At a more advanced level, it is important to establish connections between processes operating on those central processing units (CPUs) via a communication system of some kind. Distributed programming often involves one of much fundamental architecture:

- Client/server: three-tier, n-tier, or peer-to-peer. It may also be categorized as either loose coupling or tight coupling [66]. Client/server architectures include intelligent clients that communicate with a server to get data, which they then organize and present to the users. The client's input is delivered back to the server and considered a permanent modification when it is committed [67].
- Three-tier architectures include the relocation of client intelligence to an intermediate step, allowing the usage of stateless clients. This streamlines the process of deploying applications. The majority of online applications use three-tier architecture [68].
- N-tier architectures often refer to web apps that delegate their requests to other corporate services. This particular program is mostly responsible for the success of application servers [69]. The allocation of all tasks is evenly distributed across all computers, referred to as peers. Peers have the ability to function as both clients and servers. Notable instances of this architectural design include BitTorrent & the bitcoin network [70].

Another fundamental feature of distributed computing architecture is the mechanism of communicating and synchronizing tasks across concurrent processes [71]. Processes may establish direct communication with one other via different message transmission protocols, usually in a master/slave relationship [72]. On the other hand, a "database-centric" design allows distributed computing to be performed without the need for direct inter-process communication. This is achieved by employing a shared database [73]. Database-centric design, namely, enables relational processing analytics inside a structured framework, facilitating real-time data transmission in the environment [74]. This allows for the execution of distributed computing tasks, both inside and beyond the boundaries of a networked database [75].

Data mining is the systematic process of identifying and uncovering patterns in large collections of data. It utilizes techniques that combine machine learning, statistics, and database systems [76, 77]. Data mining is an area of computer science and statistics that combines many disciplines to extract information from a data collection using intelligent approaches. The ultimate objective is to turn this information into an intelligible structure that can be used for future purposes [78,79]. Data mining is the analytical phase of the "knowledge discovery in databases" process, often known as KDD. In addition to the raw analysis stage, this process also includes factors related to database and data administration, data pre-processing, model and inference considerations, interestingness measures, complexity considerations, post-processing of identified structures, visualization, and online updating [80].

The phrase "data mining" is inaccurately named, since its objective is to extract patterns and information from extensive datasets, rather than extracting the data itself [81]. Furthermore, it is commonly used as a trendy term to describe any type of extensive data or information handling, such as gathering, extracting, storing, analyzing, and using statistics. It also encompasses the use of computer systems to aid decision-making, including artificial intelligence techniques like machine learning, as well as business intelligence. The book "Data mining: Practical machine learning tools and techniques with Java" was originally going to be titled "Practical machine learning" but the phrase "data mining" was included for marketing purposes. Frequently, the broader concepts of data analysis and analytics, namely artificial intelligence in addition to machine learning techniques, are more suitable [85].

Data mining involves the analysis of large amounts of data to extract patterns that were previously unknown. These patterns can include groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining, sequential pattern mining) [86]. The process can be done either semi-automatically or automatically. Typically, this entails using database approaches like spatial indices [87]. These patterns may be regarded as a concise representation of the input data and can be used for further analysis, such as in machine learning and predictive analytics [88]. For instance, during the data mining process, it is possible to find several clusters within the data. These clusters may then be used by a decision support system to provide more precise prediction outcomes [89]. The data mining stage does not include data collection, data preparation, result interpretation, or reporting. However, these tasks are part of the entire KDD process as extra steps [90]. Data analysis and data mining differ in their purpose and methods. Data analysis is employed to evaluate models and hypotheses using a dataset, such as examining the impact of a marketing campaign, without considering the data size. On the other hand, data mining utilizes machine learning and statistical models to discover covert or concealed patterns within a substantial amount of data [91].

3. LITERATURE REVIEW

REVIEW OF PAST STUDIES:

Computation and data dispersion are really helpful in solving problems and may be used in many dispersed applications. Dispersed systems are networks that link and coordinate dispersed compute units to handle large-scale and high-performance computing requirements, which have been extensively discussed in recent decades [92]. Various types of distributed systems have been extensively researched, such as peer-to-peer (P2P) systems, grids, computational systems, ad-hoc networks, and online social network systems. Distributed networks are being used for several purposes, such as hosting web servers, storing data, and doing extensive scientific computations. Data mining has undergone extensive investigation.

Previously, conventional approaches assumed that the data is organized into clusters and stored in computer memory. This reasoning is no longer valid in the context of dispersed networks. Furthermore, the implementation of standard mining methods on remote datasets becomes unsuccessful due to the significant coordination cost involved [93]. [94] Introduced concurrent techniques for mining Correlated Hitters from a two-dimensional data stream. The researchers developed and implemented a message-passing algorithm, as well as a shared memory method, in a hybrid approach. Web crawlers were effectively integrated into major search engines, including search fields, at a time of fast expansion in Internet technology and increasing societal demands [21]. F. Liu and W. Xin [95] illustrate the integration of the Spark-based distributed crawler system architecture. They give a framework diagram that showcases the distributed framework platform in detail, leveraging Spark's RDD elastic computational model and job assignment algorithm. Nevertheless, we may address the problem of inadequate resource utilization and subpar collection efficiency by using this Spark-based distributed crawler technique. This will subsequently resolve the conflict between the current exponential increase in data volume and the pace at which information is gathered [96,97].

A proposal was made to develop an engineering system that relies on a Web crawler to create a database system of an engineering standard. This system aims to improve the efficiency of querying engineering specifications and provide a public service portal for intellectual construction drawing analysis [94]. The system primarily acquires unstructured engineering-related information from websites using the crawler module. It collects graphical information from websites using the image recognition module. The non-critical information is filtered and organized using the data cleaning module. Finally, the system ensures real-time updates of data using the data updating module.

Formal Concept Analysis (FCA) is used in several domains, such as data processing, artificial intelligence, and software engineering. The algorithms used in Formal Concept Analysis (FCA) are computationally expensive, characterized by an uneven recursion tree. In order to manage the computational complexity of the FCA, many parallel techniques have been devised [98],[20]. Therefore, it is crucial to develop a mining framework that is more scalable and adaptive in order to uncover previously unnoticed but relevant and valid patterns and information from dispersed and complicated datasets, rather of relying on centralized databases. Data mining of dispersed networks has emerged as a crucial area of study to tackle these problems [99].

Matrix decomposition is a key method used to extract information from large datasets generated by current applications. Nevertheless, the task of handling very vast quantities of data on a single computer remains undependable or unattainable. Moreover, huge data is often distributed, compiled, & kept across several devices. Consequently, such data often consist of a significant amount of diverse noise. The development of distributed matrix decomposition using big data analytics is crucial and beneficial. The distributed Bayesian matrix of decomposition model (DBMD) is introduced as a method for clustering and large-scale data mining [100,101].

An engineer responsible for the development or upkeep of a system will find value in doing study on adaptability and the capacity to evolve. The paper explores the use of evolution mining and change mining in the context of growing dispersed networks. We introduce the Service Transition Classified dependent Interface Slicing method, which extracts update information from two iterations of the distributed system of evolution. Additionally, four Metrics of Service Evolution are suggested to accurately measure the progression of the system. When the two are merged, they become the basis of our current Service Evolution Analytics approach, which includes acquiring knowledge throughout the deployment phase.

Several large-scale training methodologies have been used, prioritizing individualized design from start to finish and advanced synchronization assistance. ZenLDA is an LDA training framework specifically designed for distributed data systems [99].

The information discovered in the context of cloud computing pertains to the unauthorized access to private data by data owners. The concept of distributed databases offers a solution to this challenge by allowing several parties to divide data vertically or horizontally. Cluster analysis is a commonly used data mining technique that aims to divide a typically multivariate dataset into groups, with the goal of increasing the similarity of data items within each group [35,15,17]. This technique is particularly useful for addressing the issue of data security in distributed data mining systems.

Prior studies have shown that distributed data mining is a powerful tool for end-users, governments, & companies to analyze data and discover various types of valuable information. It broadens the possibilities of data mining, but it also presents new obstacles. Nevertheless, [92] introduced a system framework for building a comprehensive information graph of prominent academics' data. This system also identifies meta routes between items and calculates the significance of entities in the database.

Recently, [103] examined the issue of detecting extensive flexibility patterns. The widespread distribution of a vast quantity of information across many tracking sensors, both in terms of time and space, is a prevalent challenge faced by mobility tracking systems. Consequently, they create a spatial testing & information exchange protocol that provides probabilistic assurances of identifying significant patterns.

Pooled mining enables block chains to transition into clustered networks by delegating decision-making power to pool administrators [104]. While previous research has touched on this topic, the majority of these studies only offer a preliminary analysis of a particular aspect of distributed systems. For example, there is a study on grid load balancing [11,105], an article on load balancing in cloud computing, and a survey on capacity balancing in (P2P) networks.

Pooled mining allows block chains to evolve into clustered networks by entrusting decision-making authority to pool managers [104]. Although prior research has addressed this subject, the majority of these works provide just an initial examination of a certain facet of distributed systems. For instance, there have been research conducted on grid load balancing [11,105], load balancing in cloud computing, as well as capacity balancing in peer-to-peer (P2P) networks.

4. THE CHALLENGES OF DISTRIBUTED SYSTEMS

Unlike typical centralized systems, a distributed system refers to a vast array of resources that are spread out across computers linked over a network. Software sharing, hardware sharing, service sharing, and data sharing are a few illustrative instances. The proliferation of parallel computing, collaborative computing, and distributed computing has fostered the development of distributed systems. It refers to a system where networked computers communicate and synchronize their activities only via the exchange of messages [15,106].

A distributed system is a collection of autonomous computer components (subsystems) that seem to users as a unified and integrated system. The distributed system has evolved into a highly advanced system that requires cutting-edge technology and intricate algorithms, as seen in Figure 1.

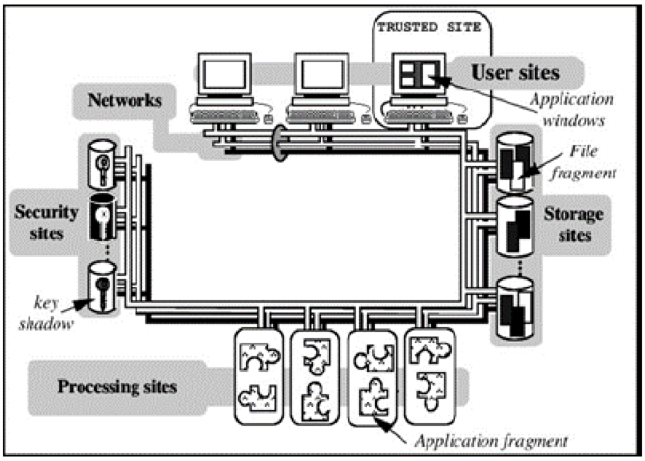


FIGURE 1: ARCHITECTURE OF DISTRIBUTED SYSTEMS

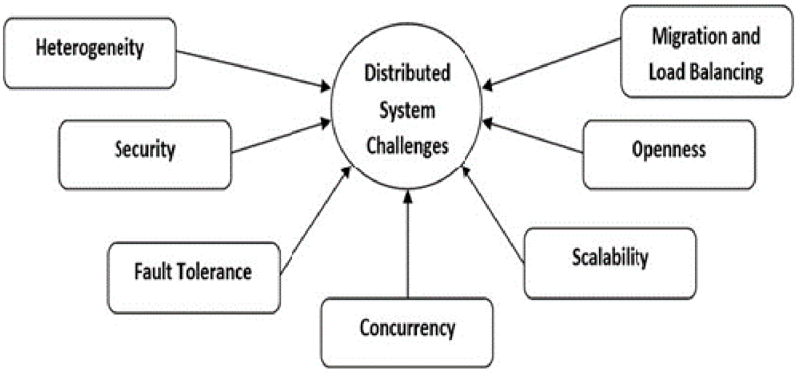


FIGURE 2: TECHNICAL DIFFICULTIES WITH THE DISTRIBUTED SYSTEM

Distributed systems, which have interconnected computational components that are spread out and organized over networks to meet the requirements of large-scale and high-speed computing, have garnered significant attention in recent decades. Various types of distributed system applications are now undergoing substantial investigation. Presently, distributed systems are used in several domains like online service, data mining, scientific computing, and data storage. Although there have been notable progressions in distributed systems, there are also some technological obstacles. In Figure 2, the primary objective of distributed systems may be classified into eight categories: heterogeneity, security, openness, failure management, scalability, concurrency, quality of service, and transparency [106].

5. CHALLENGES IN DISTRIBUTED DATA MINING

Traditional data mining methods are based on the premise that data is stored in memory, located in one place, and does not change over time [23]. The exponential increase in data volume over the last several decades has given birth to two significant concerns. Initially, an immense volume of data is produced within a little timeframe. Furthermore, the data is distributed across several locations, and consolidating it into a single site is getting progressively costly. Distributed data mining is a prominent concern in many intricate network datasets. In this distributed system, probes are strategically dispersed across the network, especially in locations with limited energy and memory [107]. Distributed data mining explores methods for implementing data mining in a decentralized fashion, using data obtained from many distributed sources. The goal is to minimize the amount of data sent between different places. The distributed data mining problem is now being explored, with a focus on minimizing communication cost, mining heterogeneous data from sources such as multisource databases, and conducting multi-relationship mining in distributed systems [14]. Figure 2 illustrates that while implementing distributed data mining, there are six technical concerns in the distributed system that remain consistent. These issues include heterogeneity, security, scalability, concurrency, quality of service, and transparency. Specifically, the issues of heterogeneity, scalability, and security are especially relevant in the context of distributed data mining.

Distributed data mining tackles these challenges in the examination of distributed data by offering several algorithmic approaches to carry out diverse mining operations and data analysis in a distributed fashion, while considering resource limitations.

The ever-increasing amount of data available from various sources poses new challenges in successfully understanding them. The process of uncovering information in large data repositories involves computationally complex, collaborative, and distributed methods and activities. The Grid is a profitable system for efficiently managing data mining and the exploration of information. In order to do this, the development of KDD applications necessitates the use of contemporary software tools and services. The KN is an advanced framework that offers tools and services for developing awareness using a grid-based approach. These services enable users to create and oversee advanced applications that integrate data sources and data mining techniques as distributed grid services, with the goal of achieving awareness. The current activities include the development and execution of services, such as WSRF-compliant Grid Services, with a new design and implementation.
The size of databases is expanding to such an extent that conventional methods of data analysis and presentation are becoming obsolete. Data mining and knowledge discovery in databases focus on extracting models and patterns of interest from large datasets (KDD). Data mining methods rely on statistical techniques, pattern analysis, and database recognition. These technologies are all examples of artificial intelligence, including high-performance computing and parallel visualization.

6. ENSURING THE CONFIDENTIALITY OF DATA IN DISTRIBUTED DATA MINING.

The Internet has emerged as the primary mode of communication for individuals and businesses worldwide. Ensuring the security of information and confidential data is crucial for the effective operation of any networking system. The cellular network is now one of the most reliable and widely used network technologies in the modern world. Shared networks have the capacity to manage large quantities of databases and data collections [14].

Proposed a technique in [14] for cellular network route intelligence mining with large data volumes. By reducing sequence data by distributed clustering and preserving anonymity, this method allows for scalability. This approach collects additional privacy protection and preservation by first aggregating raw data and then using a statistical poll. This is a well-thought-out strategy for protecting privacy and safeguarding the required database for distributed data mining [17].

A common method of maintaining privacy is to provide various nodes anonymous IDs. To preserve their anonymity, researchers have used anonymous Id assignments of different nodes in a number of research articles. In [11], they used an algorithm to distribute private information, including the anonymous assignment of IDs, across several nodes. In this investigation, they assigned 1 to N ID numbers for N nodes in order to protect privacy and security.

In networks where many nodes are responsible for exchanging data, privacy may be protected and security enhanced. For this reason, safe multiparty computing with privacy protection is a great method for exchanging and mining data. Nodes keep track of the data transfer operations, with one node handling data sharing and receiving from neighboring nodes. A node in the network that receives data employs the linear technique to be a devoted node. Using protected multiparty computing, [15] provide a method for maintaining privacy in distributed data mining.

In this study, friendly nodes for a given node that receives data in a distributed database sharing network were found using the well-known linear technique. Cloud computing provides a better foundation for sharing data and information on a common platform in today's networking systems. In the modern world, data mining has to become more well-liked among academics as a way to exchange massive volumes of data [31].
To date, a number of methods have been developed for protecting database privacy in distributed networks, or in any network. There are many uses for various hard computing techniques in privacy protection. To safeguard data in a network, soft computing methods like fuzzy logic and artificial neural networks may be used in this situation. Soft computing techniques proved to be low-cost models in a number of scenarios [108].

This technique ensures the confidentiality of individuals and network components without affecting the ultimate output of the Neural Network. While there is a possibility of data loss in this study, privacy is upheld [15]. Safeguarding one's privacy is a key consideration in any networking system. Preserving privacy and safeguarding confidential data in an impromptu network is a significant obstacle. In an ad-hoc network, nodes establish contact only when required, and this connection may be terminated after the data transfer. Once the connection is terminated, the loss of data and information may become permanent [109]. Sharing the node data and node number with nearby nodes makes it easy to identify fraudulent activity inside the network. [17] Outline a methodology to guarantee confidentiality and retain records of data loss in an ad-hoc network. This study used a system to keep track of the nodes, with nearby nodes playing a crucial role in facilitating data transmission and ensuring privacy protection.

7. DISTRIBUTED CLUSTERING WITH DYNAMIC PARALLELIZATION.

Clustering algorithms are very appealing for the purpose of detecting and extracting patterns of relevance from datasets. Nevertheless, when dealing with extensive geographical datasets, there are many obstacles that arise, including the presence of a large number of dimensions, variations in data types, and the intricate nature of some algorithms. Distributed clustering algorithms provide a very effective solution to address the issues posed by Big Data. Contemporary serial computers are significantly burdened by the considerable computing requirements of the typical Dynamic Programming (DP) methods. Optimization strategies are always intriguing when used to solve real difficulties. The minimization of a cost function is essential for achieving an optimum decision strategy that meets the specified requirements. Moreover, while dealing with practical issues, it is often essential to take into account restrictions, which serve to provide boundaries for both state and choice variables. This information adds a significant level of intricacy to the optimization issue. Combinatorial search and optimization strategies include searching for a solution to a problem from a large set of viable solutions. Due to the impracticality of exhaustive search, a directed search approach is suggested for several search and optimization issues. Furthermore, instead of only pursuing the most optimum option, there is sometimes a desire to find a satisfactory non optimal solution. Dynamic programming (DP) is a well recognized and effective approach for addressing a wide range of optimization problems, using Bellman's Principle of Optimality. There are many and well recognized uses of dynamic programming (DP). These include optimizing circuit layouts in very large-scale integration (VLSI) to reduce wire area, scheduling tasks, editing strings, packaging, managing inventories, controlling systems automatically, artificial intelligence, economics, and more. Nevertheless, this approach has not gained widespread use owing to the problem of combinatorial explosion. While it is possible to use dynamic programming analytically for some applications, in most cases, the solution must be determined numerically, and the size of the problem is a crucial factor.

The distributed clustering and dynamic parallel model are designed to overcome the limits of current distributed clustering and parallel approaches, while effectively addressing problems related to huge data. It merges the characteristics of hierarchical and partitioning techniques, and importantly, it avoids the challenge of determining the number of partitions in clustering, as well as the issues related to the termination conditions in hierarchical clustering. The total number of final clusters is constantly evaluated and formed in a hierarchical manner. Although many of these qualities show great promise, only a few have been thoroughly researched and analyzed on a small scale [21]. The distributed dynamic clustering approach consists of two distinct stages: the global model and the local model. Consequently, the act of changing local clusters by substituting nodes in the network will lead to significant additional costs and a substantial decrease in the speed of the operation. This is one of the most significant concerns associated with the majority in distributed clustering techniques. In order to address this difficulty, the objective is to minimize the amount of points that need to be exchanged between nodes.
Instead of sharing the whole data sets of all the clusters, just trade their sample points, which represent 1.2 percent of the total dataset size. The morphology and compactness of a spatial cluster are the most effective means of representing it. The boundary points of a cluster define its shape (refer to Figure 3).

Multiple strategies for detecting cluster borders may be found in different sources of literature. The form approach employs triangulation to generate the boundaries of the clusters. This approach is efficient at generating non-convex boundaries at a high speed. Distributed clustering and dynamic parallel analysis process data in a parallel and distributed way, while minimizing node communications. In the present version, the HDFS distributes data to different nodes in a random manner. The Hadoop system oversees the parallel and distributed attributes of the algorithm.

8. DISCUSSION AND COMPARISON

Every contributor has a distinct perspective, yet shares common objectives in crafting this study on the performance of distributed data mining systems. Data mining is used in distributed systems with the aim of expediting the storing, processing, analysis, and management of vast quantities of data. Every author provides a distinct description of the distributed system; what exactly is data mining? The distributed system design, as well as the data mining capabilities of a distributed system. Moreover, the performance of data mining in distributed systems compared to other standard approaches.

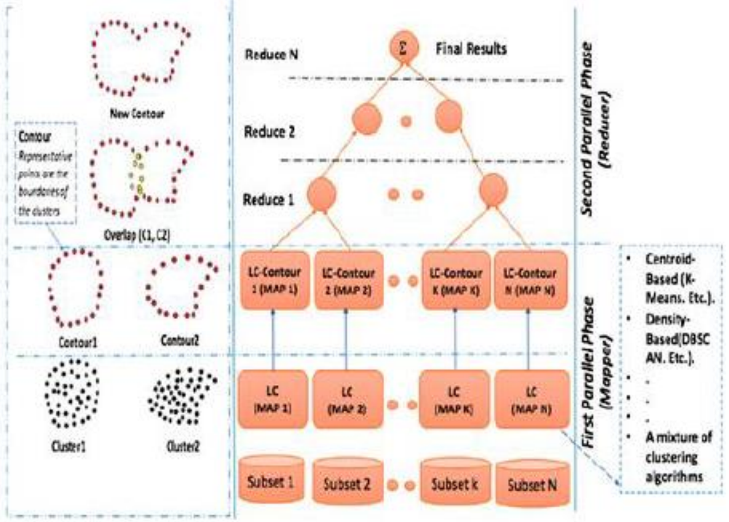
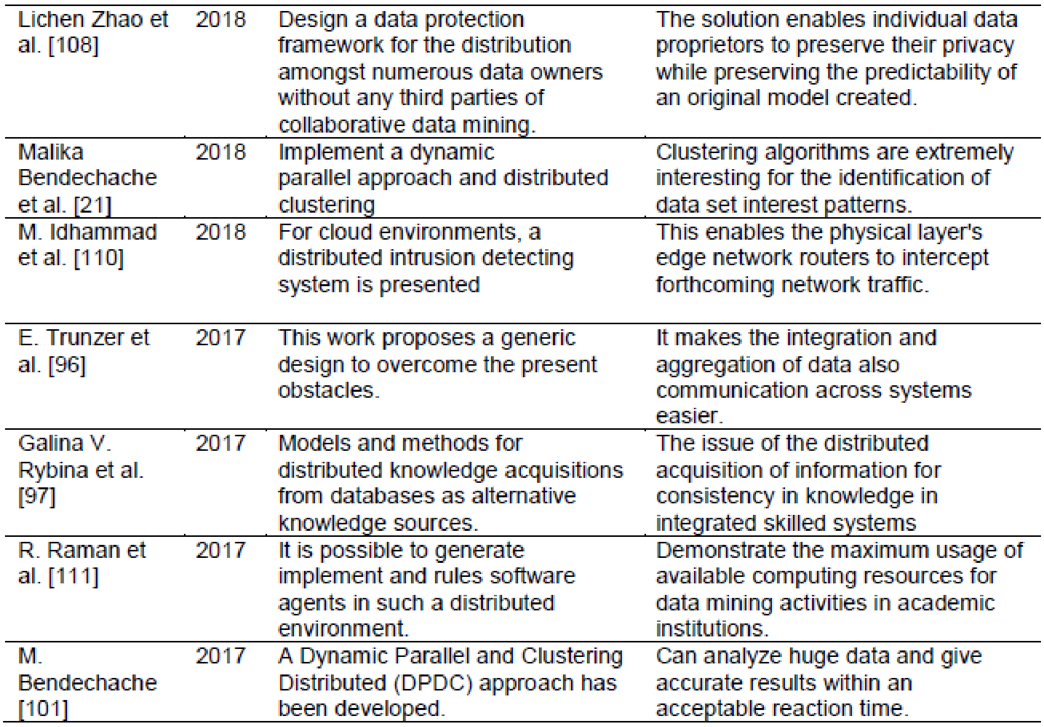


FIGURE 3 A COMPREHENSIVE PICTURE OF THE DPDC APPROACH.

TABLE 1 ASSESSMENT OF PERFORMANCE IN DISTRIBUTED SYSTEMS.

Author(s)	year	Author's objective	Description
M Vijayaraj et al. [112]	2023	Parallel and Distributed Computing for High Performance Applications	High-performance apps are able to effectively handle computationally demanding tasks thanks in large part to parallel and distributed Computing.
Donta, P.K et al. [113]	2023	Exploring the Potential of Distributed Computing Continuum Systems	Currently, distributed computing continuum systems (DCCSs) are unleashing the era of a computing paradigm that unifies various computing resources, including cloud, fog/edge computing, the Internet of Things (IoT), and mobile devices into a seamless and integrated continuum.

Author(s)	Year	Author's objectives	Description
C. Zhang et al. [100]	2020	Introduce a model of distributed Bayesian matrices decomposition for clustering and big data mining.	One of the main methods for discovering knowledge from huge data created by current applications is matrices decomposition.
F. Liu and W. Xin [92]	2020	Integrates architecture of the crawler system spark based distributed as well as a detailed introduction to distributed framework system	We can overcome the problem of inadequate resource utilization with a crawler system of spark-based distribution.
Chaturvedi et al. [95]	2020	evolution mining and change mining are used to changing distributed systems in this study.	An engineer dealing with evolution or maintenance might benefit from evolvability and changeability analysis.
Y. Shen et al. [94]	2019	implementation and Design of an engineering standardized database system dependent on data mining	To build a public service portal for intelligent construction drawing evaluation.
P. Lekshmy and M. A. Rahiman [15]	2019	It is proposed to develop a unique method-based of Privacy-Preserving of Distributed Data Mining.	The privacy-based objective variable will be constructed to expand on the privacy notion.
I. Anikin and R. Gazimov [17]	2019	Considering the issue of privacy-protected data mining in distributed networks.	Privacy-protection DBSCAN clustering algorithm protects data stored in distributed repositories.
L. Ren and P. A. Ward [104]	2019	Created and validated a Proof of performed Works' mining model.	They proposed and empirically determined the idea of equivalent blocks depending on the Poisson process model.
B. Zhao et al. [99]	2018	Present ZenLDA, an LDA training system with a generalized approach for the distributed data-parallel system.	Upon on distributed data-parallel system, the ZenLDA algorithm achieves scalable and efficient LDA training.
Da Zhang [110]	2018	Formulating a distributed software and hardware architecture to manage massive scholar data.	A system framework for constructing massive scholarly data into large knowledge graphs and discovering relationships between entities.
M. Pulimeno et al. [93]	2018	Describe parallel approaches for mining connected heavy hitters from a double data stream.	Parallel algorithms focused on the passage of the message, common memory, and hybrid mining technique.
S. Patel et al. [98]	2018	Due to computational skew, parallelizing LCM is challenging.	It makes the finding of busy workers easier and gives a technique to identify terminations.
P. Katsikouli et al. [103]	2018	Described a distributed approach for locating popular paths.	Even with a limited application of sensors and a limited user sample, fairly reliable findings may be obtained.
J. Jin et al. [31]	2018	Using HBase, electricity data is stored	How can electrical power data be controlled using HBase, Apache's distributed database



Lichen Zhao et al. [108]	2018	Design a data protection framework for the distribution amongst numerous data owners without any third parties of collaborative data mining.	The solution enables individual data proprietors to preserve their privacy while preserving the predictability of an original model created.
Malika Bendechache et al. [21]	2018	Implement a dynamic parallel approach and distributed clustering	Clustering algorithms are extremely interesting for the identification of data set interest patterns.
M. Idhammad et al. [110]	2018	For cloud environments, a distributed intrusion detecting system is presented	This enables the physical layer's edge network routers to intercept forthcoming network traffic.
E. Trunzer et al. [96]	2017	This work proposes a generic design to overcome the present obstacles.	It makes the integration and aggregation of data also communication across systems easier.
Galina V. Rybina et al. [97]	2017	Models and methods for distributed knowledge acquisitions from databases as alternative knowledge sources.	The issue of the distributed acquisition of information for consistency in knowledge in integrated skilled systems
R. Raman et al. [111]	2017	It is possible to generate implement and rules software agents in such a distributed environment.	Demonstrate the maximum usage of available computing resources for data mining activities in academic institutions.
M. Bendechache [101]	2017	A Dynamic Parallel and Clustering Distributed (DPDC) approach has been developed.	Can analyze huge data and give accurate results within an acceptable reaction time.

9. CONCLUSIONS

Data mining algorithms often aim to identify valuable patterns or carry out tasks such as classification, clustering, and outlier detection. Data analysis is often performed on applications and data that are inherently dispersed in nature. Data mining in distributed computing environments has emerged as a crucial field of study owing to the problems and difficulties associated with traditional data mining methods while handling distributed data. However, only a small number of academics have integrated the much advancement in several types of distributed data mining systems and instead produced a comprehensive classification of these systems. In this study, we analyze the definitions, overall structures, and notable characteristics of a distributed system. Our major focus is on analyzing the latest advancements in distributed data mining and offering them as our primary contributions.

REFERENCES

1. Ibrahim BR, Khalifa FM, Zeebaree SR, Othman NA, Alkhayyat A, Zebari RR, et al. "Embedded System for Eye Blink Detection Using Machine Learning Technique," in 2021 1st Babylon International Conference on Information Technology and Science (BICITS), 2021; 58-62.

2. Hasan DA, Zeebaree SR, Sadeeq MA, Shukur HM, Zebari RR, Alkhayyat AH. "Machine Learning-based Diabetic Retinopathy Early Detection and Classification Systems-A Survey," in 2021 1st Babylon International Conference on Information Technology and Science (BICITS), 2021;16-21.

3. Jijo BT, Zeebaree SR, Zebari RR, Sadeeq MA, Sallow AB, Mohsin S, et al. A comprehensive survey of 5G mm-wave technology design challenges. Asian Journal of Research in Computer Science. 2021;1-20.

4. Kareem FQ, Zeebaree SR, Dino HI, Sadeeq MA, Rashid ZN, Hasan DA, et al. A survey of optical fiber communications: challenges and processing time influences, Asian Journal of Research in Computer Science. 2021;48-58.

5. Abdullah SMSA, Ameen SYA, Sadeeq, S. Zeebaree MA. Multimodal emotion recognition using deep learning. Journal of Applied Science and Technology Trends. 2021; 2:52-58.

6. Sadeeq MA, Zeebaree S. Energy management for internet of things via distributed systems. Journal of Applied Science and Technology Trends. 2021;2:59-71.

7. Omer MA, Zeebaree SR, Sadeeq MA, Salim BW, Mohsin SX, Rashid ZN, et al. "Efficiency of malware detection in android system: A survey," Asian Journal of Research in Computer Science. 2021;59-69.

8. Maulud DH, Zeebaree SR, Jacksi K, Sadeeq MAM, Sharif KH. State of art for semantic analysis of natural language processing," Qubahan Academic Journal. 2021;1:21-28.

9. Sadeeq MM, Abdulkareem NM, Zeebaree SR, Ahmed DM, Sami AS, Zebari RR. IoT and Cloud computing issues, challenges and opportunities: A review," Qubahan Academic Journal. 2021;1:1-7.

10. Abdullah RM, Ameen SY, Ahmed DM, Kak SF, Yasin HM, Ibrahim IM, et al. Paralinguistic Speech Processing: An Overview. Asian Journal of Research in Computer Science. 2021;34-46.

11. Bagavathi A, Mummoju P, Tarnowska K, Tzacheva AA, Ras ZW. Sargs method for distributed actionable pattern mining using spark. In 2017 IEEE international conference on big data (big data), 2017; 4272-4281.

12. He P. An end-to-end log management framework for distributed systems. In 2017 IEEE 36th Symposium on Reliable Distributed Systems (SRDS). 2017;266-267.

13. Lv Z, Deng W, Zhang Z, Guo N, Yan G. A Data Fusion and Data Cleaning System for Smart Grids Big Data," in 2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), 2019;802-807.

14. Swain DK, Mishra S, Rout SB. Privacy preservation in distributed data mining for protein secondary structure prediction. In 2017 2nd International Conference on Communication and Electronics Systems (ICCES). 2017;122-127.

15. Lekshmy P, Rahiman MA. A sanitization approach for privacy preserving data mining on social distributed environment," Journal of Ambient Intelligence and Humanized Computing. 2020;11:2761-2777, 2020.

16. Ibrahim IM, Ameen SY., Yasin H. M., Omar N, Kak SF, Rashid ZN, et al., "Web Server Performance Improvement Using Dynamic Load Balancing Techniques: A Review," Asian Journal of Research in Computer Science. 2021;47-62.

17. Anikin I, Gazimov R. Approach to Privacy Preserved Data Mining in Distributed Systems," in 2019 International Russian Automation Conference (RusAutoCon), 2019;1-5.

18. Ahmed DM, Ameen SY, Omar N, Kak SF, Rashid ZN, Yasin HM, et al. A State of Art for Survey of Combined Iris and Fingerprint Recognition Systems.Asian Journal of Research in Computer Science. 2021;18-33.
19. Maulud DH, Ameen SY, Omar N, Kak SF, Rashid ZN, Yasin HM, et al. Review on Natural Language Processing Based on Different Techniques," Asian Journal of Research in Computer Science. 2021;1-17.

20. Salih AA, Ameen SY, Zeebaree SR, Sadeeq MA, Kak SF, Omar N, et al. Deep Learning Approaches for Intrusion Detection. Asian Journal of Research in Computer Science. 2021;50-64.

21. Bendechache M, Tari A-K, Kechadi M-T, Parallel and distributed clustering framework for big spatial data mining," International Journal of Parallel, Emergent and Distributed Systems. 2019;34:671-689.

22. Hassan RJ, Zeebaree SR, Ameen SY, Kak SF, Sadeeq MA, Ageed ZS, et al. State of art survey for iot effects on smart city technology: challenges, opportunities, and solutions. Asian Journal of Research in Computer Science. 2021;32-48.
23. Triantafillou P. Towards intelligent distributed data systems for scalable efficient and accurate analytics," in 2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS), 2018;1192-1202.

24. Yahia HS, Zeebaree SR, Sadeeq MA, Salim NO, Kak SF, Adel A-Z, et al. Comprehensive survey for cloud computing based nature-inspired algorithms optimization scheduling. Asian Journal of Research in Computer Science. 2021;1-16.

25. Ageed ZS, S. R. Zeebaree, M. M. Sadeeq, S. F. Kak, Z. N. Rashid, A. A. Salih, et al., "A survey of data mining implementation in smart city applications," Qubahan Academic Journal, vol. 1, pp. 91-99, 2021.

26. Ageed ZS, Zeebaree SR, Sadeeq MA, Abdulrazzaq MB, Salim BW, Salih AA, et al. A state of art survey for intelligent energy monitoring systems. Asian Journal of Research in Computer Science. 2021; 46-61.

27. Anikin IV, Gazimov RM. Privacy preserving DBSCAN clustering algorithm for vertically partitioned data in distributed systems," in 2017 International Siberian Conference on Control and Communications (SIBCON), 2017;1-4.

28. Salih A, Zeebaree ST, Ameen S, Alkhyyat A, Shukur HM. A Survey on the Role of Artificial Intelligence, Machine Learning and Deep Learning for Cybersecurity Attack Detection. in 2021 7th International Engineering Conference "Research & Innovation amid Global Pandemic"(IEC), 2021;61-66.

29. Abdullah DM, Ameen SY, Omar N, Salih AA, Ahmed DM, Kak SF, et al. Secure data transfer over internet using image steganography. Asian Journal of Research in Computer Science. 2021;33-52.

30. Kareem FQ, Ameen SY, Salih AA, Ahmed DM, Kak SF, Yasin HM, et al. SQL injection attacks prevention system technology. Asian Journal of Research in Computer Science. 2021;13-32.

31. Jin J, Song A, Gong H, Xue Y, Du M, Dong F, et al. Distributed storage system for electric power data based on hbase," Big Data Mining and Analytics. 2018;1:324-334.

32. Ismael HR, Ameen SY, Kak SF, Yasin H. M, Ibrahim IM, Ahmed AM, et al. Reliable communications for vehicular networks," Asian Journal of Research in Computer Science. 2021;33-49.

33. Abdulla AI, Abdulraheem AS, Salih AA, Sadeeq M, Ahmed AJ, Ferzor BM, et al. Internet of things and smart home security. Technol. Rep. Kansai Univ, 2020;62:2465-2476.

34. Abdulraheem AS, Salih AA, Abdulla AI, Sadeeq M, Salim N, Abdullah H, et al. Home automation system based on IoT; 2020.

35. Harikrishnasairaj K, Prasad VK. Secure frequent itemset mining from horizontally distributed databases. In 2017 International Conference on Intelligent Computing and Control (I2C2). 2017;1-4.

36. Agrawal N, Kaur A. An algorithmic approach for text recognition from printed/typed text images. in 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2018;876-879.

37. Lin C-Y, Lin Y-C. An overall approach to achieve load balancing for Hadoop Distributed File System. International Journal of Web and Grid Services. 2017; 13:448-466.

38. Salih AA, Zeebaree S, Abdulraheem AS, Zebari RR, Sadeeq M, Ahmed OM. Evolution of mobile wireless communication to 5G revolution. Technology Reports of Kansai University, 62:2139-2151.

39. Dino HI, Zeebaree S, Salih AA, Zebari RR, Ageed ZS, Shukur HM, et al. Impact of Process Execution and Physical Memory-Spaces on OS Performance," Technology Reports of Kansai University. 2020; 62:2391-2401, 2020.

40. Hamdi SJ, Ibrahim IM, Omar N, Ahmed OM, Rashid ZN, Ahmed AM, et al. A Comprehensive Study of Malware Detection in Android Operating Systems.

41. Ageed ZS, Ahmed AM, Omar N, Kak SF, IM. Ibrahim HM. Yasin, et al. A State of Art Survey of Nano Technology: Implementation, Challenges, and Future Trends.

42. Abdulqadir MM, Salih AA, Ahmed OM, Hasan DA, Haji LM, Ahmed SH, et al. A Comprehensive Study of Caching Effects on Fog Computing Performance.

43. Yazdeen AA, SR, Zeebaree MM. Sadeeq SF, Kak OM, Ahmed, Zebari RR. FPGA implementations for data encryption and decryption via concurrent and parallel computation: A review. Qubahan Academic Journal. 2021;1:8-16.

44. Ageed ZS, Zeebaree SR, Sadeeq MM, Kak SF, Yahia HS, Mahmood MR, et al. Comprehensive survey of big data mining approaches in cloud systems," Qubahan Academic Journal. 2021;1:29-38.

45. Abdulrahman LM, Zeebaree SR, Kak SF, Sadeeq MA, Adel A-Z, Salim BW, et al. A state of art for smart gateways issues and modification. Asian Journal of Research in Computer Science. 2021;1-13.

46. Abdulqadir HR, Zeebaree SR, Shukur HM, Sadeeq MM, Salim BW, Salih AA, et al., A study of moving from cloud computing to fog computing, Qubahan Academic Journal. 2021;1: 60-70.

47. AL-Zebari A, Zeebaree S, Jacksi K, Selamat A. ELMS–DPU ontology visualization with Protégé VOWL and Web VOWL," Journal of Advanced Research in Dynamic and Control Systems, 2019; 11:478-85.

48. Zeebaree A, Adel A, Jacksi K., and Selamat A. Designing an ontology of E-learning system for duhok polytechnic university using protégé OWL tool," J Adv Res Dyn Control Syst. 2019;11:24-37, 2019.

49. Adel A-Z, Zebari S, Jacksi K. Football Ontology Construction using Oriented Programming," Journal of Applied Science and Technology Trends. 2020;1:24-30.

50. SA, A-Z, Selamat A. Electronic Learning Management System Based on Semantic Web Technology: A Review," Int. J. Adv. Electron. Comput. Sci. 2017;4:1-6.

51. Abdullah RM, Abdulazeez AM, and A. Al-Zebari. Machine learning Algorithm of Intrusion Detection System. Asian Journal of Research in Computer Science, pp. 1-12, 2021.

52. Shukur H, Zeebaree SR, Ahmed AJ, Zebari RR, Ahmed O, Tahir BSA, et al. A state of art survey for concurrent computation and clustering of parallel computing for distributed systems," Journal of Applied Science and Technology Trends, 2020;1:148-154.

53. Tahir B, Ali J, Saktioto M, Fadhali R, Rahman, Ahmed A. A study of FBG sensor and electrical strain gauge for strain measurements," Journal of optoelectronics and advanced materials. 2008;10:2564-2568.

54. Harki N, Ahmed A, Haji L. CPU scheduling techniques: A review on novel approaches strategy and performance assessment. Journal of Applied Science and Technology Trends. 2020;1:48-55.

55. Ahmed A, Ahmed O. Correlation pattern among morphological and biochemical traits in relation to tillering capacity in sugarcane (Saccharum Spp). Acad J Plant Sci. 2012;5:119-122.

56. Ahmed AJ, Mohammed FH, Majedkan NA. An Evaluation Study of an E-Learning Course at the Duhok Polytechnic University: A Case Study," Journal of Cases on Information Technology (JCIT), 2022;24:1-11.

57. Ahmed O, Geraldes R, Ahmed A., DeLuca G, Palace J. Multiple sclerosis and the risk of venous thrombosis: a systematic review," in MULTIPLE SCLEROSIS JOURNAL, 2017;757-758.

58. Salim NO, Abdulazeez AM. Human diseases detection based on machine learning algorithms: A review," International Journal of Science and Business. 2021; 5:102-113.

59. Salim NO, Zeebaree SR, Sadeeq MA, Radie A, Shukur HM, Rashid ZN. Study for Food Recognition System Using Deep Learning. In Journal of Physics: Conference Series, 2021, p. 012014.

60. Salim NO, Abdulazeez AM. Science and Business," International Journal, 5;102-113.

61. Eesa AS, Sadiq S, HASSAN M, Orman Z. Rule generation based on modified cuttlefish algorithm for intrusion detection system," Uludağ University Journal of The Faculty of Engineering, vol. 26, pp. 253-268, 2021.

62. Eesa AS. Optimization Algorithms For Intrusion Detection System: A Review. International Journal of Research-Granthaalayah,2020; 8:217-225.

63. Haji SH, Zeebaree SR, Saeed RH, Ameen SY, Shukur HM, Omar N, et al., "Comparison of software defined networking with traditional networking," Asian Journal of Research in Computer Science, 2021;1-18.

64. Zebari S, Yaseen NO. Effects of parallel processing implementation on balanced load-division depending on distributed memory systems," J. Univ. Anbar Pure Sci. 2011;5:50-56.

65. Malallah H, Zeebaree SR, Zebari RR, Sadeeq MA, Ageed ZS, Ibrahim IM, et al. "A comprehensive study of kernel (issues and concepts) in different operating systems," Asian Journal of Research in Computer Science. 2021;16-31.

66. Yasin HM, Zeebaree SR, Sadeeq M. A., Ameen S. Y., Ibrahim I. M., Zebari R. R., et al. IoT and ICT based smart water management, monitoring and controlling system: A review," Asian Journal of Research in Computer Science. 2021;42-56.

67. Ibrahim I. M., Task scheduling algorithms in cloud computing: A review," Turkish Journal of Computer and Mathematics Education (TURCOMAT), 2021;12:1041-1053.

68. Zebari I. M., S. R. Zeebaree, and H. M. Yasin, "Real time video streaming from multi-source using client-server for video distribution," in 2019 4th Scientific International Conference Najaf (SICN), 2019;109-114.

69. Yasin H. M., S. R. Zeebaree, and I. M. Zebari, "Arduino based automatic irrigation system: Monitoring and SMS controlling," in 2019 4th Scientific International Conference Najaf (SICN), 2019;109-114.

70. Zeebaree S, Yasin H. M. Arduino based remote controlling for home: power saving, security and protection," International Journal of Scientific & Engineering Research, 2014;5:266-272.

71. Hasan DA, B. K. Hussan, S. R. Zeebaree, D. M. Ahmed, O. S. Kareem, and M. A. Sadeeq, "The impact of test case generation methods on the software performance: A review," International Journal of Science and Business, 2021;5:33-44.

72. Jacksi K, R. K. Ibrahim, S. R. Zeebaree, R. R. Zebari, and M. A. Sadeeq, "Clustering documents based on semantic similarity using HAC and K-mean algorithms," in 2020 International Conference on Advanced Science and Engineering (ICOASE), 2020;205-210.

73. Sadeeq MA, A. M. Abdulazeez, "Neural networks architectures design, and applications: A review," in 2020 International Conference on Advanced Science and Engineering (ICOASE), 2020; 199-204.

74. S. Zeebaree and I. Zebari, "Multilevel client/server peer-to-peer video broadcasting system," International Journal of Scientific & Engineering Research. 2014;5: 260-265.

75. Ageed Z. S., R. K. Ibrahim, and M. Sadeeq, "Unified ontology implementation of cloud computing for distributed systems," Current Journal of Applied Science and Technology, 2020;82-97.

76. Zeebaree S., S. Ameen, and M. Sadeeq, "Social media networks security threats, risks and recommendation: A case study in the kurdistan region," International Journal of Innovation, Creativity and Change 2020; 13:349-365.

77. Zebari D. A., H. Haron, S. R. Zeebaree, and D. Q. Zeebaree, "Multi-Level of DNA Encryption Technique Based on DNA Arithmetic and Biological Operations," in 2018 International Conference on Advanced Science and Engineering (ICOASE), 2018, pp. 312-317.

78. Sulaiman M. A., M. Sadeeq, A. S. Abdulraheem, and A. I. Abdulla, "Analyzation study for gamification examination fields," Technol. Rep. Kansai Univ, vol. 62, pp. 2319-2328, 2020.

79. Zeebaree S. R., A. B. Sallow, B. K. Hussan, and S. M. Ali, "Design and simulation of high-speed parallel/sequential simplified DES code breaking based on FPGA," in 2019 International Conference on Advanced Science and Engineering (ICOASE), 2019;76-81.

80. Sadeeq M., A. I. Abdulla, A. S. Abdulraheem, and Z. S. Ageed, "Impact of electronic commerce on enterprise business," Technol. Rep. Kansai Univ. 2020; 62: 2365-2378.

81. Alzakholi O., H. Shukur, R. Zebari, S. Abas, and M. Sadeeq, "Comparison among cloud technologies and cloud performance," Journal of Applied Science and Technology Trends, 2020;1:40-47.

82. Ageed Z., M. R. Mahmood, M. Sadeeq, M. B. Abdulrazzaq, and H. Dino, "Cloud computing resources impacts on heavy-load parallel processing approaches," IOSR Journal of Computer Engineering (IOSR-JCE), 2020;22:30-41.

83. Ibrahim BR, SR. Zeebaree, and B. K. Hussan, "Performance Measurement for Distributed Systems using 2TA and 3TA based on OPNET Principles," Science Journal of University of Zakho, 2019;7:65-69.

84. Sallow A., S. Zeebaree, R. Zebari, M. Mahmood, M. Abdulrazzaq, and M. Sadeeq, "Vaccine tracker," SMS reminder system: Design and implementation; 2020.

85. Sadeeq M. A., S. R. Zeebaree, R. Qashi, S. H. Ahmed, and K. Jacksi, "Internet of Things security: a survey," in 2018 International Conference on Advanced Science and Engineering (ICOASE). 2018; 162-166.

86. Abdulazeez A. M., S. R. Zeebaree, and M. A. Sadeeq, "Design and implementation of electronic student affairs system," Academic Journal of Nawroz University, 2018;7: 66-73.

87. Zeebaree S, R. R. Zebari, K. Jacksi, and D. A. Hasan, "Security approaches for integrated enterprise systems performance: A Review," Int. J. Sci. Technol. Res, 2019;8.

88. Sallow AB, M. Sadeeq, RR. Zebari, M. B. Abdulrazzaq, M. R. Mahmood, H. M. Shukur, et al., "An investigation for mobile malware behavioral and detection techniques based on android platform," IOSR Journal of Computer Engineering (IOSR-JCE), 2020;22:14-20.

89. Dino H, M. B. Abdulrazzaq, S. Zeebaree, A. B. Sallow, R. R. Zebari, H. M. Shukur, et al., "Facial expression recognition based on hybrid feature extraction techniques with different classifiers," TEST Engineering & Management. 2020;83: 22319-22329.

90. Zeebaree S, R. R. Zebari, and K. Jacksi, "Performance analysis of IIS10. 0 and Apache2 Cluster-based Web Servers under SYN DDoS Attack," TEST Engineering & Management, 2020;83: 5854-5863.

91. Jader OH, S. Zeebaree, and R. R. Zebari, "A state of art survey for web server performance measurement and load balancing mechanisms," International Journal of Scientific & Technology Research,2019;8:535-543.

92. Zhang D, Kabuka MR. Distributed relationship mining over big scholar data," IEEE Transactions on Emerging Topics in Computing, 2018; 9:354-365.

93. Pulimeno M., I. Epicoco, M. Cafaro, C. Melle, and G. Aloisio, "Parallel mining of correlated heavy hitters on distributed and shared-memory architectures," in 2018 ieee international conference on big data (big data), 2018;5111-5118.

94. Shen Y., M. Wang, H. Zhou, Q. Zhu, S. Ma, M. Cao, et al., "Design and Implementation of Engineering Standard Database System Based on Data Mining," in 2019 18th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), 2019;124-127.

95. Liu F, Xin W. Implementation of Distributed Crawler System Based on Spark for Massive Data Mining. in 2020 5th International Conference on Computer and Communication Systems (ICCCS), 2020;482-485.

96. Trunzer E, I. Kirchen, J. Folmer, G. Koltun, and B. Vogel-Heuser, "A flexible architecture for data mining from heterogeneous data sources in automated production systems," in 2017 IEEE International Conference on Industrial Technology (ICIT), 2017;1106-1111.

97. Rybina GV, Y. M. Blokhin, and E. S. Sergienko, "Distributed knowledge acquisition basing on integration of Data Mining and Text Mining methods and their usage with AT-TECHNOLOGY workbench," in 2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), 2017;1-6.

98. Patel S, Agarwal U, Kailasam S. A Dynamic Load Balancing Scheme for Distributed Formal Concept Analysis, in 2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS), 2018;489-496.

99. Zhao B, Zhou H, Li G, Huang Y. ZenLDA: Large-scale topic model training on distributed data-parallel platform," Big Data Mining and Analytics, 2018;1:57-74.

100. Zhang C, Yang Y, Zhang W, Zhang S, "Distributed Bayesian Matrix Decomposition for Big Data Mining and Clustering," arXiv preprint arXiv:2002.03703, 2020.

101. Bendechache M, Le-Khac N-A, Kechadi M-T. "Performance evaluation of a distributed clustering approach for spatial datasets," in Australasian Conference on Data Mining, 2017;38-56.

102. Chaturvedi A, Tiwari A, Binkley D, Chaturvedi S. Service evolution analytics: change and evolution mining of a distributed system," IEEE Transactions on Engineering Management. 2020;68:137-148.

103. Katsikouli P, Astefanoaei MS, Sarkar R, "Distributed mining of popular paths in road networks," in 2018 14th International Conference on Distributed Computing in Sensor Systems (DCOSS), 2018;1-8.

104. Ren L, Ward PA. Pooled mining is driving blockchains toward centralized systems," in 2019 38th International Symposium on Reliable Distributed Systems Workshops (SRDSW), 2019;43-48.

105. Carrillo GE, Abad, CL. Inferring Workflows with Job Dependencies from Distributed Processing Systems Logs," in 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on
Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), 2017;1025-1030.

106. Mole PV. Empirical Research on the Challenges of Distributed Databases: A Literature Review."

107. Rana MS, Sohel MK, Arman MS. "Distributed Database Problems Approaches and Solutions-A Study," in International Journal of Machine Learning and Computing (IJMLC), ed; 2018.

108. Zhao L, Ni L, Hu S, Chen Y, Zhou P, Xiao FF, et al. "Inprivate digging: Enabling tree-based distributed data mining with differential privacy," in IEEE INFOCOM 2018-IEEE Conference on Computer Communications, 2018;2087-2095.

109. Rathnayake RS, Poravi G. Review on Textual Data Mining for Reviewer Recommendation in Pull-Based Distributed Software Development," in 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 2019;1-5.

110. Idhammad M, Afdel K, Belouch M. "Distributed intrusion detection system for cloud environments based on data mining techniques," Procedia Computer Science, 2018;127:35-41.

111. Raman R, Vadivel S, Raj BE. A framework for cost-effective distributed data mining in academic institutions using intelligent agents," in 2017 Fourth HCT Information Technology Trends (ITT), 2017;13-18.

112. M Vijayaraj,R.Malar Vizhi, P.Chandrakala, Laith H. Alzubaidi, Khasanov Muzaffar,R Senthilkumar, Parallel and Distributed Computing for High Performance Applications, E3S Web of Conferences 399, 04039 (2023) https://doi.org/10.1051/e3sconf/202339904039 ICONNECT-2023.

113. Donta, P.K.; Murturi, I.; Casamayor Pujol, V.; Sedlak, B.; Dustdar, S. Exploring the Potential ofDistributed Computing Continuum Systems. Computers 2023, 12, 198. https://doi.org/10.3390/computers12100198.