Article

Not peer-reviewed version

A Survey on Dependency of Parallel Clustering Platforms on Clustering Algorithms with Their Clustering Criteria for Big Data

Maradana Durga Venkata Prasad * and Srikanth Thota

Posted Date: 15 December 2023

doi: 10.20944/preprints202312.1201.v1

Keywords: Clustering; Machine learning; Clustering Algorithms; Clustering Criteria; Clustering Platform; Clustering Criteria; Big Data; horizontal scaling platforms; MapReduce; Spark; P2P; vertical scaling platforms; Multi-cores; GPU; FPGA

 doi:10.20944/preprints202312.1201.v1

Article

A Survey on Dependency of Parallel Clustering Platforms on Clustering Algorithms with Their Clustering Criteria for Big Data

Maradana Durga Venkata Prasad [1,*] and Dr. Srikanth T [2]

[1] Research Scholar, Department of Computer Science and Engineering, Gandhi Institute of Technology and Management (GITAM), Visakhapatnam, Andhra Pradesh, India
[2] Associate Professor, Department of Computer Science and Engineering, Gandhi Institute of Technology and Management (GITAM), Visakhapatnam, Andhra Pradesh, India. sthota@gitam.edu
* Correspondence: powersamudra@gmail.com

Abstract: Clustering is a data mining task used for the data extraction from the data bases or files. Clustering is used to find unknown groups present in the data sources like files or data bases. This paper focuses on clustering algorithms performance dependency on the parallel clustering platforms and the clustering algorithms along with their clustering criteria. The problems with the present Traditional clustering algorithms were throughput and data source size changes (scalability). So they can't address the big data. So for handling the huge volumes of data, parallel clustering algorithms along with clustering criteria were used. For processing the big Data Parallel clustering algorithms are of two types based on computing platforms used. They were the horizontal scaling platforms and vertical scaling platforms.

Keywords: clustering; machine learning; clustering algorithms; clustering criteria; clustering platform; clustering criteria; big data; horizontal scaling platforms; MapReduce; Spark; P2P; vertical scaling platforms; multi-cores; GPU; FPGA

---

1. INTRODUCTION

In the present era data analysis technique clustering used to handle the emerging challenges related to big data. Data analysis technique is applied on the data set to partition it into two subsets. One set consists of similar instances and other consists of dissimilar instances [1]. For partitioning various clustering methods were used like Bi-clustering, Density Based, Graph Based, Grid Based, Hard Clustering, Hierarchical, Model Based, Partitioning, and Soft Clustering e.t.c.

Clustering technique is used to group data points into clusters from a file or data base. i.e Similar points are grouped into one cluster and other data points to another group. Clustering purpose is to identify the similar and dissimilar characteristics or patterns from the given data. Similar points are identified using similarity functions. After the clustering process, class labels were assigned to the clusters called as classification. Clustering process takes the input from the data sources and gives the clustered data as output. Clustering used in different applications like pattern recognition, Image Processing and analysis, document categorization, Stock market segmentation, exploratory data analysis, World Wide Web, metrology, Health Care department, social network analysis e.t.c [2].The stages of clustering are shown in the Figure 1 below.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 December 2023 doi:10.20944/preprints202312.1201.v1

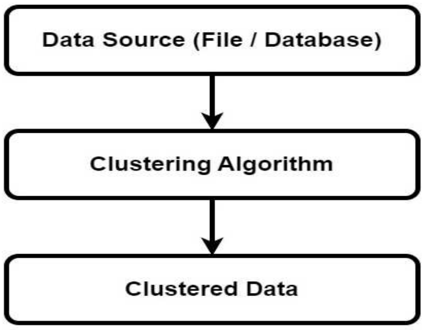


Figure 1. Clustering Stages.

Machine learning techniques are of three types. They were supervised, Semi-Supervised Learning and unsupervised. Clustering is unsupervised learning technique [3]. Machine learning techniques Classification are show in the Table 1 below.

Table 1. Machine learning techniques Classification.

Algorithms of Machine Learning List	**Sub Methods**	**Details**
Supervised Learning	Regression	It is used to predict continuous numeric values. Examples: Linear regression and support vector regression [4].
	Classification	It is used to assign data points to predefined categories. Examples: decision trees, logistic regression, random forests, support vector machines [5].
Unsupervised Learning	Clustering	It is used to group similar data points into. Examples: partition clustering, K-Means and DBSCAN [6].
	Dimensionality Reduction	It is used to remove unnecessary features from a given data set. Examples: Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE)[7].
Semi-Supervised Learning		It combines the aspects of supervised unsupervised learning [8].

1.1. Challenges

The Challenges of Traditional Clustering Algorithms are addressed using Parallel Clustering Algorithms are show in the below Table 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 December 2023 doi:10.20944/preprints202312.1201.v1

Table 2. Criteria of clustering in Parallel Clustering Algorithms.

Clustering Algorithms		
Evaluation Criteria	**Traditional**	**Parallel**
Knowledge	Require prior knowledge	Not require prior knowledge
Data	Should be ordered	Ordered not required
Input Parameters	complex	Simple
Data Set	Not partitioned	Partitioned (chunks)
Problem	Specific Problem	All Problems
Operation Conditions	particular	All Conditions
Computational Costs	More	Less
Data it handle	can't handle heterogeneous data	can handle heterogeneous data
Execution	Serial	Parallel
Speed	Less	More
Throughput	Less	More
Scalability	Less	More
Big Data Challenges	Can't Meet	Can Meet
Volume / Data Quantity (Created / Stored)	More	More

Velocity / Frequency **(Coming / Updated)**	Less	Less
Types Of Data **(Data Forms And Sources From Where It Is Coming)**	Less	More

1.2. Scope of the article

This paper focuses on parallel clustering algorithms on different Parallel computing platforms such as horizontal and the vertical scaling platforms to handle the huge volumes of Data. A horizontal scaling platform contains peer networks, MapReduce, and Spark platforms and which is used to add or remove machines to which the work load is distributed. vertical scaling platforms contains high Performance Computing Clusters (HPC), Multicore processors, Graphics Processing Unit (GPU), and Field Programmable Gate Arrays (FPGA) platforms which are used to add or remove power (processors, RAM, and hardware) to the present machine. It consists of.

1.3. Contributions

This Section deals with the contexts like different clustering algorithms with clustering criteria and parallel clustering platforms for handling huge volumes of Data.

2. LITERATURE SURVEY / Organization

This survey consists of two sections. They were

a. Study of different clustering algorithms with clustering criteria.
b. Study of different parallel clustering platforms.
c. Use of clustering platforms and clustering algorithms with clustering criteria for clustering.

3. Study of different clustering algorithms with clustering criteria

In the market Different types clustering methods were there proposed by different researcher's persons. For each clustering method there will be one or more sub clustering Algorithms. Each sub clustering algorithm will have its own constraints. The major clustering methods available in the market were shown in Table 3 below, Clustering Algorithms Dependency on Clustering criteria is shown in Figure 2 and Big data platforms is shown in Figure 3.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 December 2023 doi:10.20944/preprints202312.1201.v1

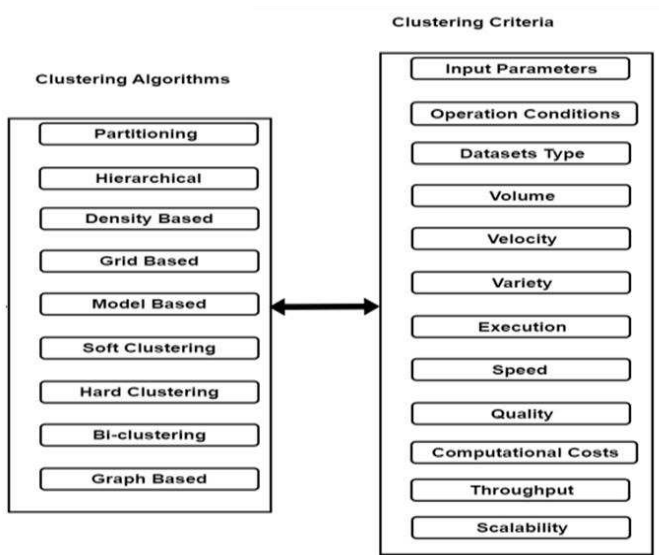


Figure 2. Clustering Algorithms Dependency on Clustering criteria.

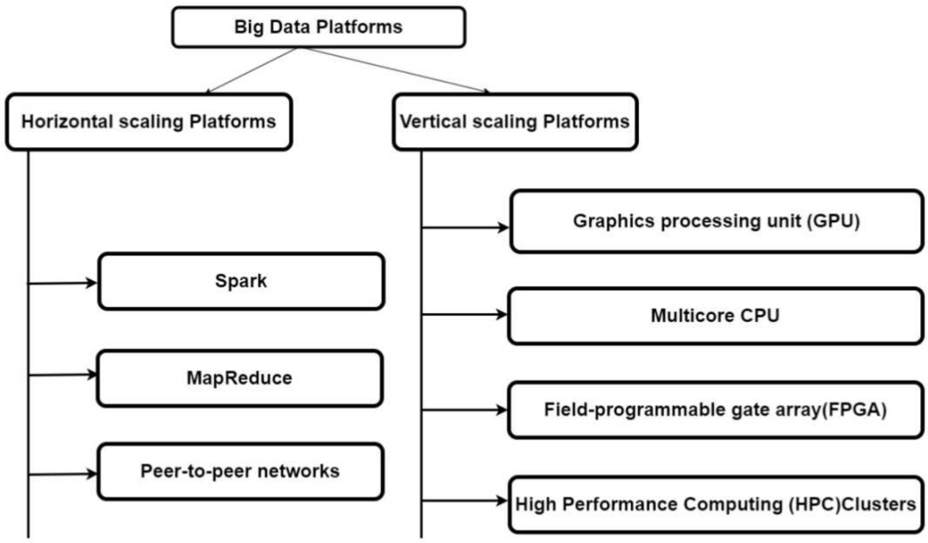


Figure 3. Big Data Platforms.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 December 2023 doi:10.20944/preprints202312.1201.v1

Table 3. Different Types of Clustering Algorithms and their sub Clustering Methods.

Clustering Algorithm	Details	Sub Clustering Methods
Partitioning	It is a technique used to break a data source into two groups [9].	1. CLARA. 2. CLARANS. 3.EMCLUSTERING 4. FCM. 5. K MODES. 6. KMEANS. 7. KMEDOIDS. 8. PAM. 9. XMEANS
Hierarchical	It creates clusters based on objects similarity [10]	1. AGNES. 2. BIRCH. 3. CHAMELEON. 4. CURE. 5. DIANA. 6. ECHIDNA 7. ROCK.
Density Based	It creates clusters based on radius as a condition. i.e within radius one cluster and remaining other cluster(noise) [11].	1. DBCLASD. 2. DBSCAN. 3. DENCLUE. 4. OPTICS.
Grid Based	Clustering is done based on calculation values of density of cells using the grid [12].	1. CLIQUE. 2. OPT GRID. 3. STING. 4. WAVE CLUSTER.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 December 2023 doi:10.20944/preprints202312.1201.v1

Model Based	It uses statistical approach of assigning weights to every object. Based on object weights clustering is done [13].	1. EM. 2. COBWEB. 3. SOMS.
Soft Clustering	It is based on assigned of individual data points to more than one cluster [14].	1. FCM. 2. GK. 3. SOM. 4. GA Clustering
Hard Clustering	It is based on assigned of individual data points to everyone cluster [15].	1. KMEANS
Bi-clustering	It creates clusters based on cluster matrix rows and columns as a condition [16].	1. OPSM. 2. Samba 3. JSa
Graph Based	It is based on graph theory that graph contains vertices or nodes. Here every node is assigned particular weights. Based on graph node weights Clustering is done [17].	1. Graph based k-means algorithm

Clustering Types:

Clustering technique is used to divide the given data set into two groups based on the objects similarity present in the data set. The Hard and Soft based Clustering groups of clustering are show in the Table 4 below and Hard and Soft based Clustering sets representation is also shown in Figure 4.

Table 4. Hard Clustering verses Soft Clustering.

Clustering Type	**Hard Clustering**	**Soft Clustering**
All Data Point Assigned to	single cluster	multiple clusters
Similarity Clustering	maximum	minimum

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 December 2023 doi:10.20944/preprints202312.1201.v1

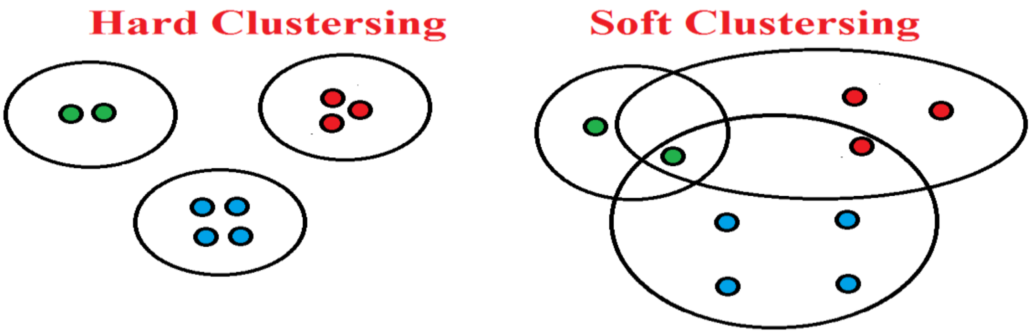


Figure 4. Soft and Hard Based Clustering.

Different Clustering Algorithms with Clustering Criteria

The clustering algorithm performance every time is based on the following constraints, parameters and user preferences. Different Clustering Algorithms with Clustering Criteria are show in the Table 5 below.

Table 5. Different Clustering Algorithms with Clustering Criteria.

Clustering Criteria	**Details**
Data Mining Tasks	It is of two types. They were Descriptive or Predictive. Clustering is Descriptive Data Mining Tasks. Descriptive Data Mining task gives us the provide correlation, cross-tabulation, frequency, etc., from the data [18]. Predictive Data Mining task is used to analyze and predict future occurrences of events or other data or trends [19].
Type of Learning / Knowledge	Machine learning is of three types. They were Supervised / Unsupervised, Reinforced learning and unsupervised. Supervised / Reinforced machine learning is based on output training data and the labeled input and [20] and unsupervised learning Machine learning processes unlabelled data [21]. Reinforcement technique is used to train algorithms to learn from their environments [22].
Dimensionality	If the clustering algorithm deals with more types of data then it is said to be multi dimensional. (High / Low / Medium) [23].
Data Sources	Data Set / File / Data Base
Volume	Number of data points of a dataset. (Created / Stored)

 doi:10.20944/preprints202312.1201.v1

<table>
<tr><td>Unstructured or Structured Data</td><td>If the data in the data set is in a standardized format (clearly defined) for the easy access by the systems or humans is called as Structured data [24]. If the data in the data set is not a standardized format (clearly defined) for the easy access by the systems or humans is called as Unstructured data [25]. Structured data is easily made into clusters but not Unstructured data. So algorithms are used to convert unstructured data to Structured data. So there is a requirement of unstructured data to be converted into unstructured data and it can discover new patterns. Clustering uses Structured in most cases.</td></tr>
<tr><td>Data Types used in Clustering</td><td>Two types of data are processed by the Clustering algorithms. They were Qualitative and Quantitative Data [26]. Clustering algorithm processes two types of data are shown in Figure 5 shown below.

Qualitative / Categorical type (Subjective) of data can be split into categories. Example: Persons Gender (male, female, or others) [27]. It is of three types. They were Nominal (sequenced), Ordinal (ordered) and binary (take true (1) / false (0)).

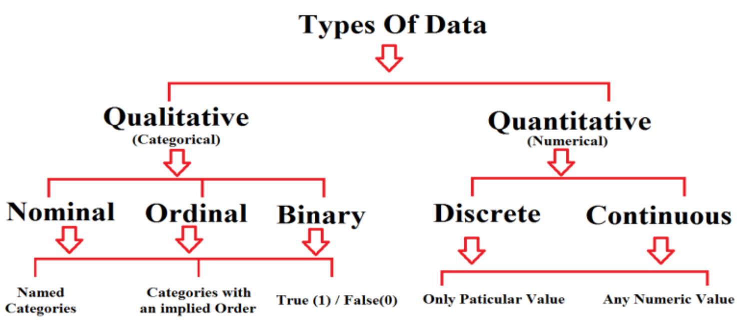
Quantitative / Numerical Data are measurable and are of two types. They were Discrete (countable, continuously, measurable) [28]. Example: Student height.


Figure 5: Clustering algorithm processes two types of data</td></tr>
<tr><td>ETL Operations used</td><td>Extraction, Transformation and loading operations are performed on the data source [29].</td></tr>
</table>

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 December 2023 doi:10.20944/preprints202312.1201.v1

Data Preprocessed	It is used for data cleaning and data transforming to make it used for the analysis [30].
Data Preprocessing Methods	Data Preprocessing Methods used in the market are cleaning, instance selection, normalization, scaling, feature selection, one-hot encoding, data transformation, feature extraction and feature selection and dimensionality reduction [31].
Hierarchical Clustering Algorithms Type	In Hierarchical clustering algorithms [32] is two types Divisive (Top-Down) [33] Or Agglomerative (Bottom-Up) [34].
No Of Clustering Algorithms	It is the total count of two types of Clustering Algorithms (Main and sub).i.e. It is count of sum of total number of Main Clustering Algorithms and total number of Sub Clustering Algorithms.
Algorithms Threshold / Stops At What Level	Hierarchical clustering algorithms Stops at a level defined by the user as his Preferences.
Algorithm Stability	It uses different clustering applications to determine the number of clusters.
Programming Language	It used For processing (Python, Java, .Net e.t.c) the clustering algorithm.
Number Of Inputs For The Clustering Process	Clustering Algorithm, Algorithm Constraints, Number of Levels and clusters per each level.

 doi:10.20944/preprints202312.1201.v1

Number Of Levels	In Hierarchical clustering algorithms, divisive clustering (top-down) how many split it goes down is the number levels. Or Agglomerative (bottom-up) how many merges it goes up to the number of levels.
Clusters Level Wise	It is number of clusters at each level or stage
Data Points per Cluster	It is always depends on the type of cluster algorithm used and its preferences defined by the user.
Similarity Functions / Similarity Measure.	It is used to quantify how similar or dissimilar two clusters are in a clustering analysis. Similarity measures are used to identify the good clusters in the given data set. There are so many Similarity measures used in the current market. They were Weighted, Average, Chord, Mahalanobis, Canberra Metric, Czekanowski Coefficient, Index of Association, Mean Character Difference, Pearson coefficient, Minkowski Metric, Manhattan or City blocks distance, KullbackLeibler Divergence, Clustering coefficient, Cosine, Kmean e.t.c[35].
Intra Cluster Distance	It is the distance between the data points of one cluster to other. If its value is low then the clusters are said to be tightly coupled other clusters are said to be loosely coupled [36].
Inter Cluster Distance	It measures the dissimilarity / separation between different clusters. It quantifies how distinct or well-separated the clusters are from each other [37].
Sum Of Square Error (SSE) Or Other Errors	It is a measure of difference the actual to the expected result of the model [38].
Clusters Likelihood	It is clusters similarity in the data points [39].
Clusters Unlikelihood	It is clusters dissimilarity in the data points.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 December 2023 doi:10.20944/preprints202312.1201.v1

Number Of Variable Parameters At Each Level	These are the input parameters which are changed during the running of the algorithm like threshold.
Outlier	In the clustering process any object doesn't belong to any cluster it is called as an outlier.
Clusters Compactness	It deals with the inertia for better clustering. It means lower inertia indicates better clustering. Inertia means Within-Cluster Sum of Squares.
Purpose	Develop and predict model
Clustering Scalability	It is the increasing and decreasing abilities of every cluster as a part o whole.
Total Number of Clusters	It is total number clusters generated by the clustering algorithm after its execution.
Interpretability	Understandability , usability of clusters after is generation is called as Interpretability
Convergence	Convergence criterion is a condition by which controls the change in cluster centers. It should be always to be minimum.
Clusters Shape	Each clustering Algorithm handles the clustering in different shapes [40]. Clustering Algorithm ------- Cluster Shape K Means ------- Hyper Spherical, Centroid Based Approach ------- Concave Shaped Clusters, Cure ------- Arbitrary, Partitional Clustering ------- Ellipsoidal, Clarans ------- Polygon Shaped, Dbscan ------- Concave E.t.c
Execution	Running the clustering Algorithm or Algorithms in serial or parallel.
Output	Clusters
Velocity	It is nothing but the frequency of Coming / Updated data to the clusters.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 December 2023 doi:10.20944/preprints202312.1201.v1

Throughput	It is the count of number of units processed by the system in the given amount of time.
Space Complexity	It of a clustering algorithm refers to the amount of memory or storage for storing input data, data structures or variables required by the algorithm to perform clustering on a given dataset. Space Complexity=Auxiliary Space + Space For Input Values.
Time Complexity	It is the time taken to run each and instructions of an algorithm. Time Complexities of Clustering Algorithms Clustering Algorithm ---- Time Complexity BIRCH ---- O(n) Chameleon ---- O(n^2) CLARA ---- O(n) CLARANS ---- O(n^2) Clique ---- O(n) CURE ---- O(s^2*s) K -Means ---- O(n) K-medoids ---- O(n^2) PAM ---- O(n^2) ROCK ---- O(n^3) Sting ---- O(n) e.t.c
Clusters Visualization	It is a process used to representing clusters or groups of data points in a visual format. It gives the insights into patterns, relationships, and structures within the data. Techniques and tools for visualizing clusters: Scatter Plots, Dendrogram, Heatmaps, t-Distributed Stochastic Neighbor Embedding, Silhouette Plots, Principal Component Analysis Plot, K-Means Clustering Plot, Hierarchical Clustering Dendrogram, Density-Based Clustering Visualization, Interactive Visualization Tools: Matplotlib, Seaborn, Plotly, D3.js, and Tableau [41].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 December 2023 doi:10.20944/preprints202312.1201.v1

4. Overview of different Parallel clustering platforms

This section gives an outline about the two different Parallel data processing platforms of big data [42]. They were

1. Horizontal scaling platforms.
2. Vertical scaling platforms.

The different Parallel clustering platforms are shown in Figure 6 and Types of Big Data Platform along with Clustering Algorithm is depicted in Figure 7 below.

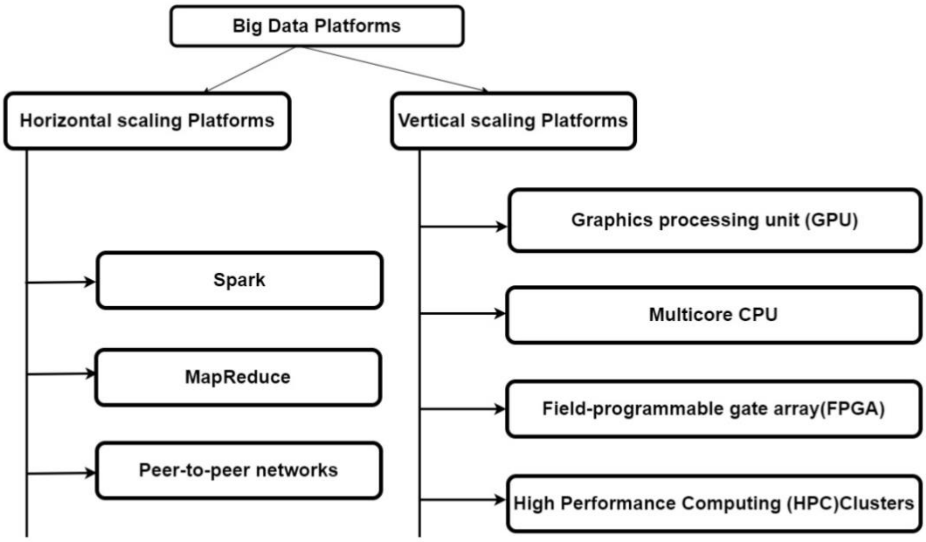


Figure 6. Types of Big Data Platforms.

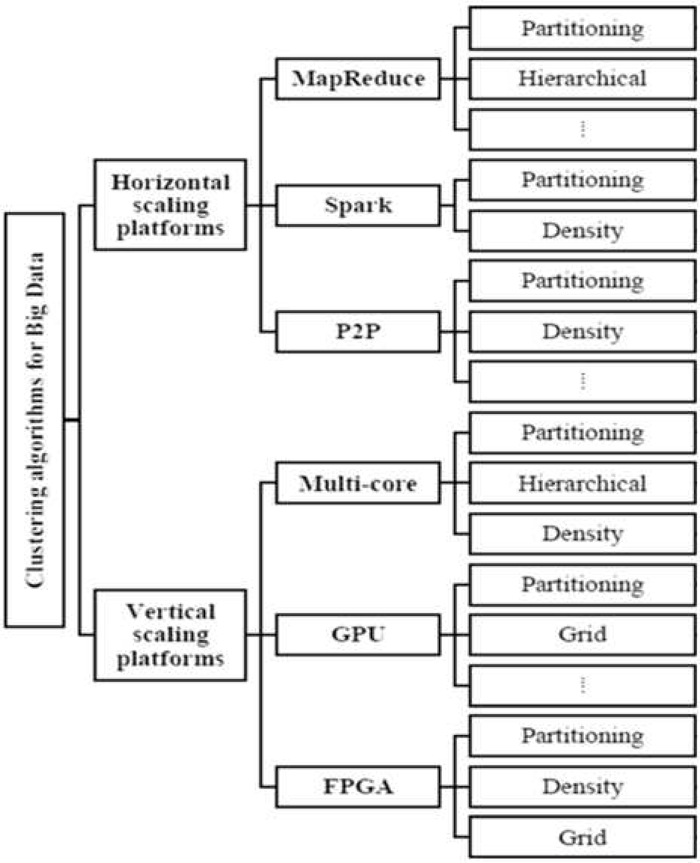


Figure 7. Types of Big Data Platform along with Clustering Algorithm.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 December 2023 doi:10.20944/preprints202312.1201.v1

Parallel clustering platform: Horizontal scaling platforms

A horizontal scaling platform contains Map Reduce, Peer-to-peer networks and Spark and vertical scaling platforms which contains Field Programmable Gate Arrays, Graphics Processing Unit and Multi-core CPU. Parallel clustering platforms we use different clustering methods like Partitioning, Hierarchical, Density Based, Grid Based, Model Based, Soft Clustering, Hard Clustering, Bi-clustering, Graph Based e.t.c.

4.1. MapReduce

It is a software framework model for parallel programming introduced by Google used for parallel processing by splitting big datasets into parts chunks and executing them in parallel on multiple commodity servers and at the end it aggregates all the data from the multiple servers and returns the output back to the application. Hadoop platform written in languages like C++, Java, Python and Ruby, Which executes the Map Reduce Programs. In cloud computing platform, Map Reduce will run in parallel suing multiple systems in the cluster for data analysis. MapReduce program work in two phases. They were Map and Reduce [43]. MapReduce phases are show in the Table 6 below.

Table 6. MapReduce phases.

MapReduce phases Name	**Phase Purpose**
Map tasks (Splits & Mapping)	splitting and mapping of data
Reduce tasks (Shuffling, Reducing)	Shuffle and reduce the data.

MapReduce program Phases

MapReduce for parallel programming model goes through the following Phases splitting, mapping, shuffling, and reducing. Normally user gets the input from the files or data base for these phases.All the Phases of MapReduce are shown in Table 7 below and All Phases of MapReduce are shown in the Figure 8 below.

Table 7. Phases of MapReduce.

Phases of MapReduce	**Details**
Splitting	The Input data set is broken down into small parts called as data chunks and is used by a single map.
Mapping	Splitting is input phase for mapping phase where each data chunk is given to the mapping function to measure the number of occurrences of each word. Mapping phase Output: (word, frequency)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 December 2023 doi:10.20944/preprints202312.1201.v1

Shuffling & Sorting	This phase is used to combine similar words with their frequency.
Reducing	It is used to give the consolidated summary of the given dataset.

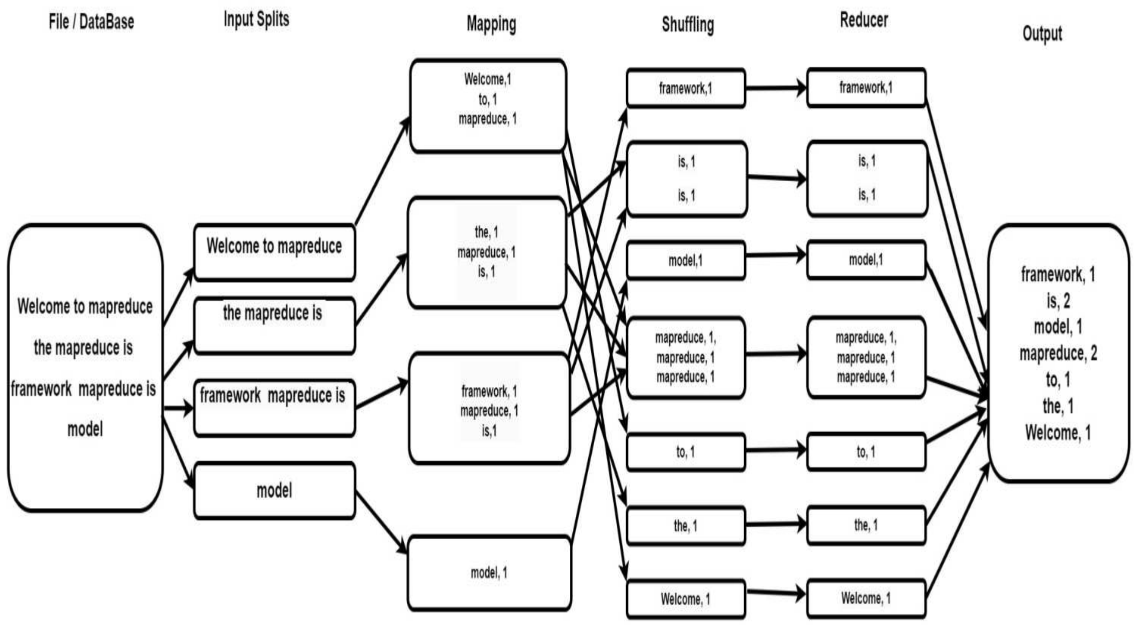


Figure 8. Phases of MapReduce.

MapReduce Work Organization in Hadoop

Jobs are divided into tasks by Hadoop. Tasks are of two types (Map tasks, Reduce tasks) were shown in Table 8 below.

Table 8. MapReduce Tasks.

MapReduce Tasks	**Purpose**
Map tasks	Splits & Mapping
Reduce tasks	Shuffling, Reducing

Execution Process of MapReduce Programs

Execution Process of MapReduce Programs is controlled by two main components (Job tracker, Multiple Task Trackers). User interacts with the Jobtracker for the completion of their job. MapReduce Execution Process Components are show in Table 9 below and MapReduce Execution Process Components is shown in Figure 9 below.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 December 2023 doi:10.20944/preprints202312.1201.v1

Table 9. MapReduce Execution Process Components.

MapReduce Execution Process Components	Details
Jobtracker	It is a master node which is used to complete execution of submitted job. Jobtracker resides on Namenode.
Multiple Task Trackers	Multiple Task Trackers are the slave machines for performing the job submitted by Jobtracker node. Task Trackers resides on Datanode

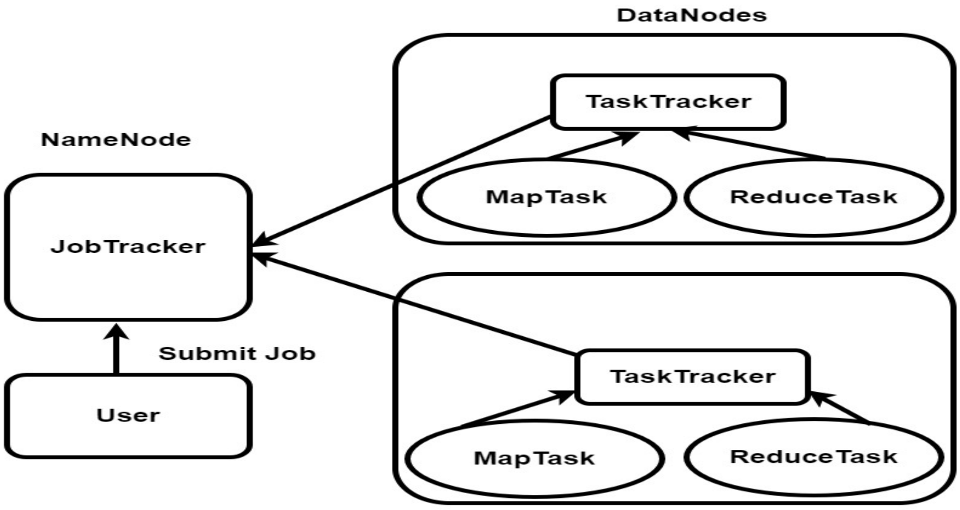


Figure 9. MapReduce Execution Process Components.

4.2. Spark Apache

Spark is used for the data processing framework for parallel processing. Apache Spark consists of two main components. They were single master node (process) and several worker nodes.

Master node assigning tasks Worker nodes and controls the Worker nodes and the resources assigned to the Worker nodes. Apache Spark runs queries, continuous iterative jobs and reduces execution time on big datasets. Apache Spark supports shared variables, resilient distributed datasets (RDDs) and parallel operations. Spark executes the applications as separate sets of processes on a cluster. Apache Spark runs the Driver Program it creates a SparkContext which is used to convert the user program / written code into jobs which runs on the cluster. Spark Driver coordinates the cluster manager to control the jobs execution. Normally jobs are broken down into sub jobs and are distributed to worker nodes. Spark executor runs the jobs and data and cache. Spark executors are registered with the spark context. Spark executor are dynamically created and removed during the running of the tasks [44]. Apache Spark main components are shown in the Figure 10 below.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 December 2023 doi:10.20944/preprints202312.1201.v1

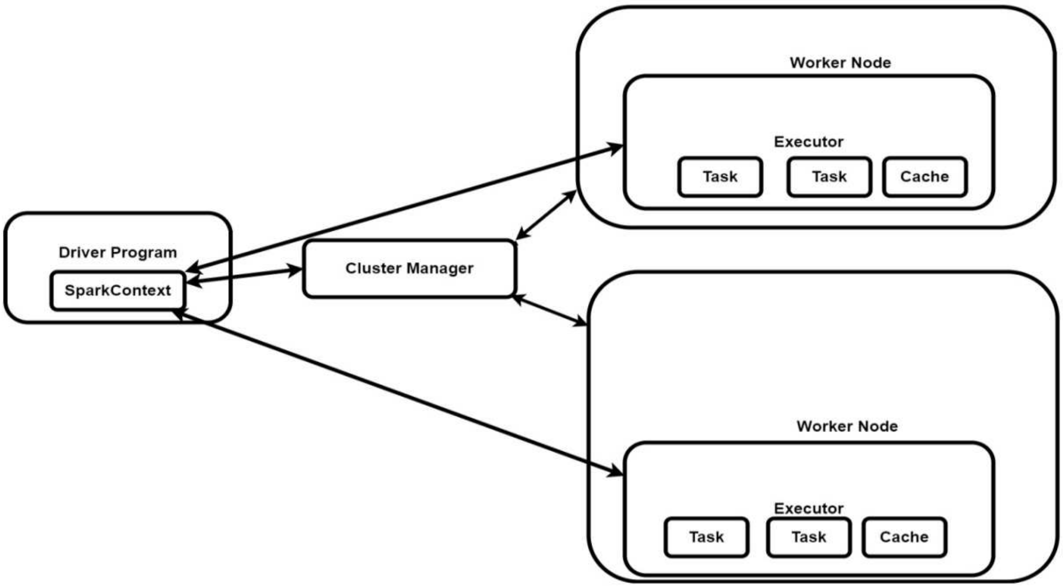


Figure 10. Apache Spark main components.

Spark Apache Execution Modes

Spark Apache runs in three different modes (Cluster mode, Client mode and Local mode) based on where your application resources are located and you're going to run. Spark Apache Execution Modes are shown in the Table 10 below.

Table 10. Spark Apache Execution Modes.

Execution Modes	**Details**
Cluster	It is used to run production jobs where the driver executes under the worker nodes.
Client	Here the driver runs locally from where you are submitting your application using spark-submit command.
Local	It is used to run complete Spark Application on a individual machine. Local mode uses threads instead of parallelized threads.

Cluster Manager Types

The cluster managers supported by the current system which were shown in the Table 11 below and Apache Spark Features are shown in Table 12.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 December 2023 doi:10.20944/preprints202312.1201.v1

Table 11. Types of Cluster Managers.

Cluster Manager Types	Details
Standalone	It is used to set up a cluster and provides a web-based graphical user interface to monitor the cluster.
Apache Mesos	It is used to run multiple distributed applications on the same cluster resource allocation and scheduling conflicts.
Hadoop YARN	It is a Hadoop3 resource manager.
Kubernetes	It an open source system for automatic scaling, management , deployment applications.

Table 12. Apache Spark Features.

Apache Spark Features	Details
Speed	Applications run on Spark process which is much faster in memory and on disk by reducing number of read and write operations to disk.
Multi-Language Support	Spark uses various APIs like Java, Scala, or Python. So application programs can be written different languages.
Advanced Analytics	For generating analytics spark uses Map, reduce, SQL queries machine learning (ML), and graph algorithms and streaming data

4.3. P2P networks

P2P means Peer-to-peer networks. P2P architecture distributes the divides tasks or workloads among peers. P2P networks not requires server to control the operations (transfer and receiving and data). All the peers are given equal preference in the p2p architecture. P2p networks are used to share files and access to devices among the peers. Nodes of the peer to peer networks can be scalable (Added / Removed). P2P networks are resistant to failures mean one node fails it effects the other nodes. Security wise P2P network are failed and require high bandwidth usage for the data transfer [45]. Typical Peer-to-peer network architecture is shown in the Figure 11 below.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 December 2023 doi:10.20944/preprints202312.1201.v1

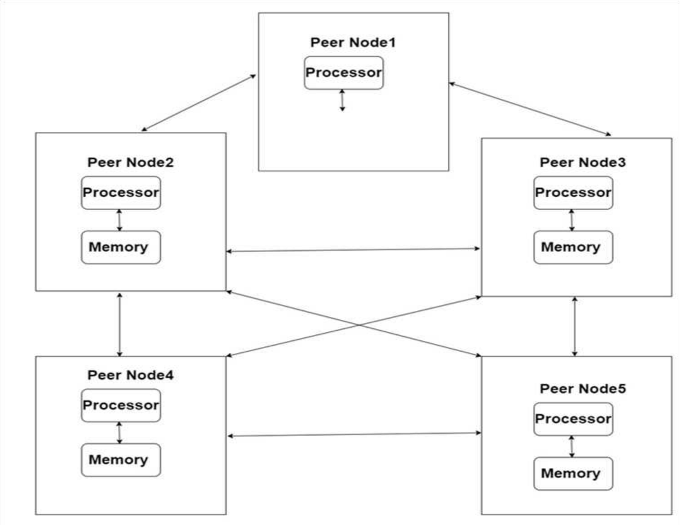


Figure 11. Typical Peer-to-peer network architecture.

Parallel clustering platform: Vertical scaling platforms

4.4. Graphics processing unit (GPU)

It is a graphics card / single chip microprocessor which is used for processing 2D and 3D Graphics. GPU uses parallel architecture as Single Instruction, Multiple Data (SIMD. Normally the CPU used to process the instruction serially known processing. But the GPU is used for vector processing and parallel computing means the processes are executed simultaneously or parallel. So parallel processing is very fast and efficient. GPUs takes a big problem breaks it into parts called as tasks and work on all these tasks and consolidate it to gets the result for the big problem. GPU parallel computing platform used to do one or more computations or processes are carried out parallel. So GPUs are fast and efficient [46]. Comparison of CPU and GPU are shown in Table 13 below. Graphics processing unit Instruction pooling is shown in the Figure 12 below.

Table 13. Comparison of CPU and GPU.

	CPU(Central processing unit)	**Graphics processing unit (GPU)**
No Of Cores	4 to 8	100s Or 1000s
Throughput	Low	High
Instructions Execution	Serial	Parallel
Computing Applications	General purpose	High Performance
Parallel Programming Languages	Java, .net e.t.c	CUDA, Opencl E.T.C

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 December 2023 doi:10.20944/preprints202312.1201.v1

Drawback		Limited Memory Capacity
Memory Management	Easy	Complex

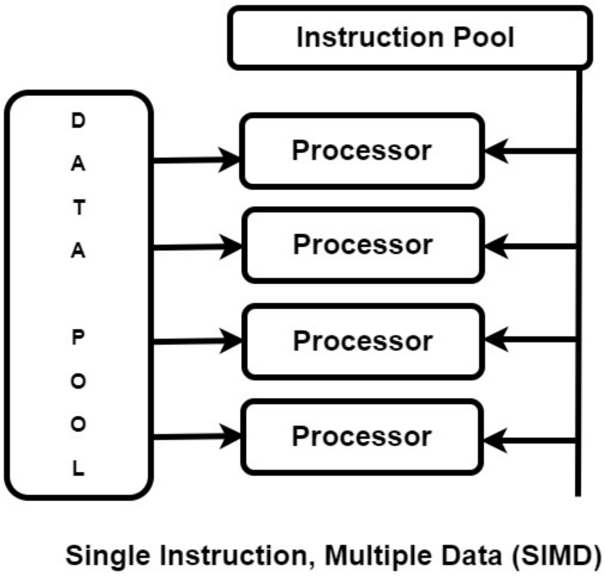


Figure 12. Typical Peer-to-peer network architecture.

5.5. Multi-core CPU

It is a single chip an integrated circuit which consists of two or more processor cores. It is used to process of multiple tasks using one of the concepts like multithreading and parallel processing. Multi cores CPU are used due to the increase in performance; reduce power consumption, low heat generation [47]. Typical Multi-core CPU is shown in the Figure 13 below.

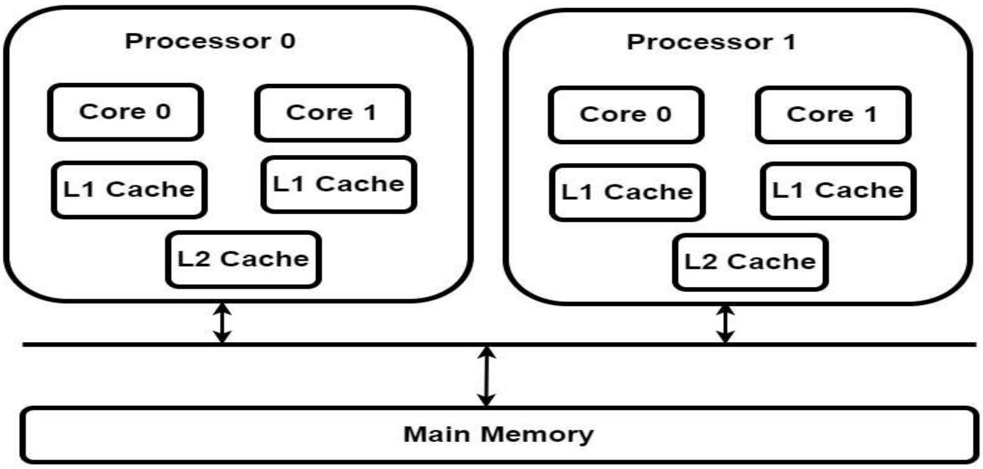


Figure 13. Typical Multi-core CPU.

6.6. Field programmable gate arrays (FPGA)

It is a integrated circuit which can be programmed for doing customized operations for a particular application. FPGA is low cost, flexible, expandable larger and less power consuming [48].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 December 2023 doi:10.20944/preprints202312.1201.v1

FPGA consists of three main structures. The FPGA structures were shown in the Table 14 below. Typical FPGA architecture is shown in the Figure 14 below.

Table 14. FPGA structure.

FPGA structure	Details
Programmable logic structure	It consists of collection of CLBs. CLB means configurable logic block used to implement any Boolean function of 4 to 6 variables. One or two flip-flops are used to implement one CLB.
Programmable routing structure	It is used for routing the information. It consists of vertical and horizontal routing channels, connection boxes and switch boxes.
Programmable Input / Output structure	It consists of buffers either which are used as Input buffers or used as output buffers.

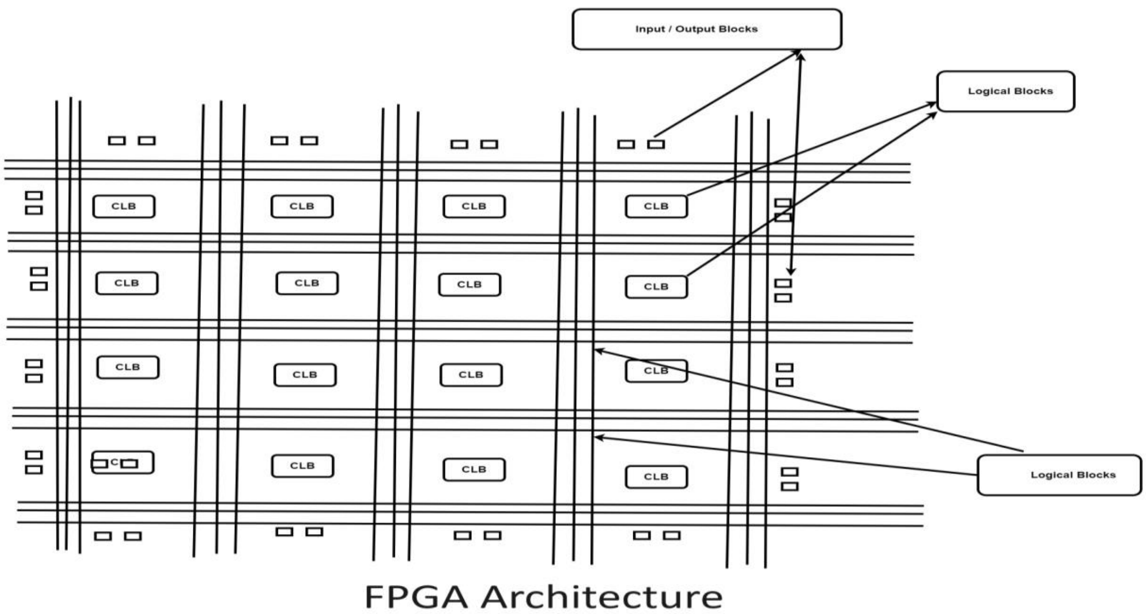


Figure 14. Typical FPGA architecture.

4. Clustering Dependence on parallel clustering platform, clustering Algorithms and clustering Criteria

The research carried out in the context of this survey of parallel clustering platforms, clustering algorithms and clustering algorithms criteria. Clustering algorithms always depends on the clustering criteria and parallel clustering platforms always depends on the clustering algorithms.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 December 2023 doi:10.20944/preprints202312.1201.v1

Dependency of parallel clustering algorithms on clustering algorithms and clustering Criteria is shown in the Table 15 below.

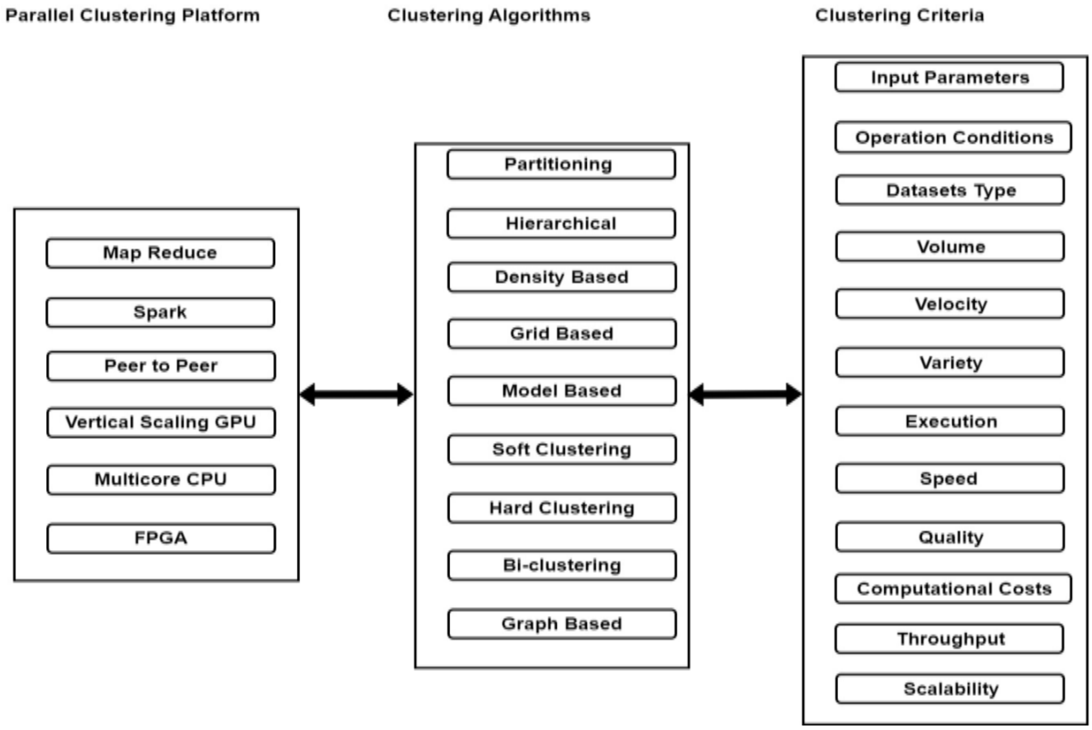


Figure 15. Dependency of parallel clustering algorithms on clustering algorithms and clustering Criteria.

Note:

1. Every algorithm uses its own data type to get optimal clusters or results [49].
2. Based on patterns, clusters, iterations and Levels Generated, clustering algorithm time and space Complexity of the clustering algorithm will varies [50].
3. Clustering method performance based on Data source, Data source size, shape of clusters shape, objective function, and similarity measurement functions [51].
4. Clustering methods use different data types like Numerical, categorical, Textual data, Multimedia, Network, Uncertain, Time Series, Discrete data e.t.c [52].
5. Similarity functions are used for identify inter and intra clusters similarities in between the clusters. Examples of distance functions are Euclidean Distance Function, Manhattan Distance Function, Chebyshev Distance Function, Davies Bould in Index e.t.c. Distance Function can affect the Performance of the clustering Algorithms [53].
6. Clustering algorithm is one of the steps in Knowledge Discovery in Databases (KDD) process [54].
7. In the clustering process Uniqueness may or may not be present in the Inter and Intra clustering process [55].
8. In any Clustering Algorithm used to differentiate between one cluster group with other cluster group [56].
9. Each and every Clustering method will have its own advantages and disadvantages based on the constraints, metrics used in the clustering algorithm [57].
10. Each clustering algorithm will have its own sub methods [58].
11. Parallel clustering platform is used to run the clustering algorithms in parallel [59].
12. Parallel clustering platform depends on clustering algorithm and clustering algorithm depends on clustering criteria [60].

 doi:10.20944/preprints202312.1201.v1

5. Conclusions

This paper is about the performance clustering algorithms always depends on the parallel clustering platforms and the clustering algorithms along with their clustering criteria. Clustering platforms are used for data processing either by using horizontal scaling platforms or vertical scaling platforms. Parallel clustering platforms use MapReduce, Peer-to-Peer networks and Spark. A vertical scaling platform forms with the Multi-core processors, GPU, and FPGA.

Conflicts of Interest: The authors declare that there are no conflicts of interest in connection with the work submitted.

References

1. Kamalpreet Bindra and Anuranjan Mishra, "A detailed study of clustering algorithms", 2017 6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) , DOI: 10.1109/ICRITO.2017.8342454.
2. Jelili Oyelade, Itunuoluwa Isewon, Olufunke Oladipupo, Onyeka Emebo, Zacchaeus Omogbadegun, Olufemi Aromolaran, Efosa Uwoghiren, Damilare Olaniyan and Obembe Olawole, "Data Clustering: Algorithms and Its Applications", DOI: 10.1109/ICCSA.2019.000-1.
3. O. Obulesu, M. Mahendra and M. ThrilokReddy, "Machine Learning Techniques and Tools: A Survey", 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), DOI: 10.1109/ICIRCA.2018.8597302.
4. Diyana Kinaneva, Georgi Hristov, Petko Kyuchukov, Georgi Georgiev, Plamen Zahariev and Rosen Daskalov, "Machine Learning Algorithms for Regression Analysis and Predictions of Numerical Data", 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) , DOI: 10.1109/HORA52670.2021.9461298.
5. Shovan Chowdhury and Marco P. Schoen, "Research Paper Classification using Supervised Machine Learning Techniques", 2020 Intermountain Engineering, Technology and Computing (IETC), DOI: 10.1109/IETC47856.2020.9249211.
6. Ritesh C. Sonawane and Hitendra D. Patil, "Clustering Techniques and Research Challenages in Machine Learning", 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), DOI: 10.1109/ICCMC48092.2020.ICCMC-00054.
7. Deddy Jobson and Tirucherai Gopalakrishnan Venkatesh, "Dimensionality Reduction Techniques to Aid Parallelization of Machine Learning Algorithms", 2022 IEEE 7th International conference for Convergence in Technology (I2CT), DOI: 10.1109/I2CT54291.2022.9825239.
8. Nitin Namdeo Pise and Parag Kulkarni, "A Survey of Semi-Supervised Learning Methods",2008 International Conference on Computational Intelligence and Security, DOI: 10.1109/CIS.2008.204.
9. A. Dharmarajan and T. Velmurugan, "Applications of partition based clustering algorithms: A survey", 2013 IEEE International Conference on Computational Intelligence and Computing Research, DOI: 10.1109/ICCIC.2013.6724235.
10. Zahra Nazari, Dongshik Kang, M. Reza Asharif, Yulwan Sung and Seiji Ogawa, "A new hierarchical clustering algorithm", 2015 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), DOI: 10.1109/ICIIBMS.2015.7439517.
11. Pradeep Singh and Prateek A. Meshram, "Survey of density based clustering algorithms and its variants", 2017 International Conference on Inventive Computing and Informatics (ICICI), DOI: 10.1109/ICICI.2017.8365272.
12. Qiang Zhang, "A Grid Based Clustering Algorithm", 2010 6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM), DOI: 10.1109/WICOM.2010.5600140.
13. H.-P. Kriegel, P. Kroger, A. Pryakhin and M. Schubert, "Effective and efficient distributed model-based clustering", Fifth IEEE International Conference on Data Mining (ICDM'05), DOI: 10.1109/ICDM.2005.53.
14. S. Rajathi, N. Shajunisha and S.Shiny Caroline, "Correlative analysis of soft clustering algorithms", 2013 Fifth International Conference on Advanced Computing (ICoAC), DOI: 10.1109/ICoAC.2013.6921977.
15. Mohiuddin Ahmed and Abu Barkat, "Performance Analysis of Hard Clustering Techniques for Big IoT Data Analytics", 2019 Cybersecurity and Cyberforensics Conference (CCC), DOI: 10.1109/CCC.2019.000-8.
16. S.C. Madeira and A.L. Oliveira, "Biclustering algorithms for biological data analysis: a survey", IEEE/ACM Transactions on Computational Biology and Bioinformatics, DOI: 10.1109/TCBB.2004.2.
17. Shimei Jin, Wei Chen and Jiarui Han, "Graph-based machine learning algorithm with application in data mining", 2017 Third International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN) , DOI: 10.1109/ICRCICN.2017.8234519.

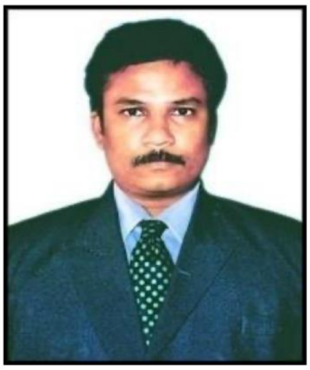
18. H. Ahonen, O. Heinonen, M. Klemettinen and A.I. Verkamo, "Applying data mining techniques for descriptive phrase extraction in digital document collections", DOI: 10.1109/ADL.1998.670374.
19. Hina Gulati, M. Klemettinen and A.I. Verkamo, "Predictive analytics using data mining technique", 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom).
20. Mohammed Amine El Mrabet, Khalid El Makkaoui and Ahmed Faize, "Supervised Machine Learning: A Survey ", 2021 4th International Conference on Advanced Communication Technologies and Networking (CommNet), DOI: 10.1109/CommNet52204.2021.9641998.
21. Nagdev Amruthnath and Tarun Gupta, "A research study on unsupervised machine learning algorithms for early fault detection in predictive maintenance ", 2018 5th International Conference on Industrial Engineering and Applications (ICIEA), DOI: 10.1109/IEA.2018.8387124.
22. Muddasar Naeem, Syed Tahir Hussain Rizvi and Antonio Coronato, "A Gentle Introduction to Reinforcement Learning and its Application in Different Fields", IEEE Access, DOI: 10.1109/ACCESS.2020.3038605.
23. Ayush Soni, Akhtar Rasool, Aditya Dubey and Nilay Khare, "Data Mining based Dimensionality Reduction Techniques", "2022 International Conference for Advancement in Technology (ICONAT)", DOI: 10.1109/ICONAT53423.2022.9725846.
24. Seema Maitrey and C. K. Jha, "Handling Structured Data Using Data Mining Clustering Techniques", 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), DOI: 10.1109/ICICT46931.2019.8977647.
25. Suyash Mishra and Anuranjan Misra, "Structured and Unstructured Big Data Analytics", " 2017 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), DOI: 10.1109/CTCEEC.2017.8454999.
26. Mike Simpson, Simon Woodman, Hugo Hiden, Sebastian Stein, Stephen Dowsland, Mark Turner, Vicki L. Hanson and Paul Watson, "A Platform for the Analysis of Qualitative and Quantitative Data about the Built Environment and Its Users", " 2017 IEEE 13th International Conference on e-Science (e-Science)", DOI: 10.1109/eScience.2017.36.
27. Dao Lam, Mingzhen Wei And Donald Wunsch, "Clustering Data of Mixed Categorical and Numerical Type With Unsupervised Feature Learning", IEEE Access, DOI: 10.1109/ACCESS.2015.2477216.
28. Avishek Bose; Arslan Munir; Neda Shabani, "A Quantitative Analysis of Big Data Clustering Algorithms for Market Segmentation in Hospitality Industry", 2020 IEEE International Conference on Consumer Electronics (ICCE DOI: 10.1109/ICCE46568.2020.9043023.
29. Rahmadi Wijaya and Bambang Pudjoatmodjo, "An overview and implementation of extraction-transformation-loading (ETL) process in data warehouse (Case study: Department of agriculture)", 2015 3rd International Conference on Information and Communication Technology (ICoICT), DOI: 10.1109/ICoICT.2015.7231399.
30. Sanjay Kumar Dwivedi and Bhupesh Rawat, "A review paper on data preprocessing: A critical phase in web usage mining process", 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), DOI: 10.1109/ICGCIoT.2015.7380517.
31. Asma Saleem, Khadim Hussain Asif, Ahmad Ali, Shahid Mahmood Awan and Mohammed A. Alghamdi, "Pre-processing Methods of Data Mining", "2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing, DOI: 10.1109/UCC.2014.57.
32. Sakshi Patel, Shivani Sihmar and Aman Jatain, "A study of hierarchical clustering algorithms", 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom).
33. Nurcan Yuruk, Mutlu Mete, Xiaowei Xu, Thomas A. J. Schweiger, "A Divisive Hierarchical Structural Clustering Algorithm for Networks", Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007), DOI: 10.1109/ICDMW.2007.73.
34. Hussain Abu Dalbouh and Norita Md. Norwawi, "Improvement on Agglomerative Hierarchical Clustering Algorithm Based on Tree Data Structure with Bidirectional Approach", 2012 Third International Conference on Intelligent Systems Modelling and Simulation, DOI: 10.1109/ISMS.2012.13.
35. Usha Rani and Shashank Sahu, "Comparison of clustering techniques for measuring similarity in articles", 2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT), DOI: 10.1109/CIACT.2017.7977377.
36. Sarra Ben Hariz and Zied Elouedi, "IK-BKM: An incremental clustering approach based on intra-cluster distance", ACS/IEEE International Conference on Computer Systems and Applications - AICCSA 2010, DOI: 10.1109/AICCSA.2010.5587008.
37. Xin Xin and Aggelos K. Katsaggelos, "A novel image retrieval framework exploring inter cluster distance", 2010 IEEE International Conference on Image Processing, DOI: 10.1109/ICIP.2010.5651817.

38. José Ortiz-Bejar, Eric S. Tellez, Mario Graff, Jesús Ortiz-Bejar, Jaime Cerda Jacobo and Alejandro Zamora-Mendez, "Performance Analysis of K-Means Seeding Algorithms", 2019 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC), DOI: 10.1109/ROPEC48299.2019.9057044.
39. R.M. Castro, M.J. Coates and R.D. Nowak, "Likelihood based hierarchical clustering", IEEE Transactions on Signal Processing (Volume: 52, Issue: 8, August 2004), DOIDOI: 10.1109/TSP.2004.831124.
40. Sajad Shirali-Shahreza, Soheil Hassas Yeganeh, Hassan Abolhassani and Jafar Habibi, "Circluster: Storing cluster shapes for clustering", IEEE Transactions on Signal Processing", 2008 4th International IEEE Conference Intelligent Systems, DOI: 10.1109/IS.2008.4670502.
41. Wenbo Wang; Yuwei Li; Feng Wang; Xiaopei Liu; Youyi Zheng", Using Visualization to improve Clustering Analysis on Heterogeneous Information Network", IEEE Transactions on Signal Processing", 2018 22nd International Conference Information Visualisation (IV), DOI: 10.1109/iV.2018.00046.
42. Hadjir Zemmouri, Said Labed and Akram Kout ", "A survey of parallel clustering algorithms based on vertical scaling platforms for big data", 2022 4th International Conference on Pattern Analysis and Intelligent Systems (PAIS), DOI: 10.1109/PAIS56586.2022.9946663.
43. Sukhwant kour Siledar, Bhagyashree Deogaonkar, Nutan Panpatte and Jayshri Pagare, "Map Reduce Overview and Functionality", 2021 6th International Conference on Communication and Electronics Systems (ICCES), DOI: 10.1109/ICCES51350.2021.9489056.
44. Eman Shaikh, Iman Mohiuddin, Yasmeen Alufaisan and Irum Nahvi, "Apache Spark: A Big Data Processing Engine", 2019 2nd IEEE Middle East and North Africa COMMunications Conference (MENACOMM), DOI: 10.1109/MENACOMM46666.2019.8988541.
45. Mirza Abdur Razzaq, Javed Ahmed Mahar, Muneer Ahmad, Najia Saher, Arif Mehmood And Gyu Sang Choi, "Hybrid Auto-Scaled Service-Cloud-Based Predictive Workload Modeling and Analysis for Smart Campus System", Digital Object Identifier 10.1109/ACCESS.2021.3065597
46. Ahmed Hussein Ali and Mahmood Zaki Abdullah, "A Survey on Vertical and Horizontal Scaling Platforms for Big Data Analytics", International Journal of Integrated Engineering • September 2019.
47. Xiaohong Qiu, Geoffrey Fox, Huapeng Yuan, Seung-Hee Bae, George Chrysanthakopoulos and Henrik Nielsen, "Parallel Data Mining on Multicore Clusters", 2008 Seventh International Conference on Grid and Cooperative Computing, DOI: 10.1109/GCC.2008.100.
48. Ahmed Hussein Ali and Mahmood Zaki Abdullah, "A Survey on Vertical and Horizontal Scaling Platforms for Big Data Analytics", https://doi.org/10.30880/ijie.2019.11.06.015.
49. Punyaban Patel and Borra Sivaiah; Riyam Patel, "Approaches for finding Optimal Number of Clusters using K-Means and Agglomerative Hierarchical Clustering Techniques", DOI: 10.1109/ICICCSP53532.2022.9862439.
50. Punyaban Patel and Borra Sivaiah; Riyam Patel, "Approaches for finding Optimal Number of Clusters using K-Means and Agglomerative Hierarchical Clustering Techniques", DOI: 10.1109/ICICCSP53532.2022.9862439.
51. Simone Santiniand Ramesh Jain, "Similarity Measures", IEEE Transactions On Pattern Analysis And Machine Intelligence.
52. Dao Lam, Mingzhen Wei And Donald Wunsch, "Clustering Data of Mixed Categorical and Numerical Type With Unsupervised Feature Learning", DOI: 10.1109/ACCESS.2015.2477216.
53. Naveen Kumar, Sanjay Kumar Yadav and Divakar Singh Yadav, "Similarity Measure Approaches Applied in Text Document Clustering for Information Retrieval", 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), DOI: 10.1109/PDGC50313.2020.9315851.
54. U. Fayyad, "Data mining and knowledge discovery in databases: implications for scientific databases", Proceedings. Ninth International Conference on Scientific and Statistical Database Management (Cat. No.97TB100150), DOI: 10.1109/SSDM.1997.621141.
55. Sachin Paranjape, S. Barani, Mukul Sutaone, Prachi Mukherji, "Intra and inter cluster congestion control technique for mobile wireless sensor networks", 2016 Conference on Advances in Signal Processing (CASP), DOI: 10.1109/CASP.2016.7746213.
56. Huican Zhu, Hong Tang and Tao Yang, "Demand-driven service differentiation in cluster-based network servers", Proceedings IEEE INFOCOM 2001. Conference on Computer Communications. Twentieth Annual Joint Conference of the IEEE Computer and Communications Society (Cat. No.01CH37213 DOI: 10.1109/INFCOM.2001.916256.
57. Nuwan Ganganath, Chi-Tsun Cheng and Chi K. Tse, "Data Clustering with Cluster Size Constraints Using a Modified K-Means Algorithm", 2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, DOI: 10.1109/CyberC.2014.36.

 doi:10.20944/preprints202312.1201.v1

58. Ritesh C. Sonawane and Hitendra D. Patil, “Clustering Techniques and Research Challenages in Machine Learning”, 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), DOI: 10.1109/ICCMC48092.2020.ICCMC-00054.
59. Bing Zhou, Jun-Yi Shen and Qin-Ke Peng, “PARCLE: a parallel clustering algorithm for cluster system”, Proceedings of the 2003 International Conference on Machine Learning and Cybernetics, DOI: 10.1109/ICMLC.2003.1264431.
60. Xia Daoping and Zhong Alin and Long Yubo, “A Parallel Clustering Algorithm Implementation Based on Apache Mahout”, 2016 Sixth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC), DOI: 10.1109/IMCCC.2016.9.

AUTHOR DETAILS:

	Dr. Srikanth Thota received his Ph.D in Computer Science Engineering for his research work in Collaborative Filtering based Recommender Systems from J.N.T.U, Kakinada. He received M.Tech. Degree in Computer Science and Technology from Andhra University. He is presently working as an Associate Professor in the department of Computer Science and Engineering, School of Technology, GITAM University, Visakhapatnam, Andhra Pradesh, India. His areas of interest include Machine learning, Artificial intelligence, Data Mining, Recommender Systems, Soft computing.

 doi:10.20944/preprints202312.1201.v1

	Mr. Maradana Durga Venkata Prasad received his B.TECH (Computer Science and Information Technology) in 2008 from JNTU, Hyderabad and M.Tech. (Software Engineering) in 2010 from Jawaharlal Nehru Technological University, Kakinada, He is a Research Scholar with Regd No: 1260316406 in the department of Computer Science and Engineering, Gandhi Institute Of Technology And Management (GITAM) Visakhapatnam, Andhra Pradesh, INDIA. His Research interests include Clustering in Data Mining, Big Data Analytics, and Artificial Intelligence. He is currently working as an Assistant Professor in Department of Computer Science Engineering, CMR Institute of Technology, Ranga Reddy, India.