International Conference on Emerging Trends in Science and Technology (ETIST 2021)
Jointly Organized by Department of Biological Science, Physical Science and Computational Science
Nallamuthu Gounder Mahalingam College, Affiliated to Bharathiar University, Tamilnadu, India.
Published by NGMC - November 2021

A Survey on Big Data in Data Mining Techniques

Dr P. Logeswari[1] , J. GokulaPriya[2], G.Banupriya[3] , S.Sudha[4], S.Sharmila[5]

@ngmc2021

Abstract

Big data processing emerges as AN innovative and potential analysis space for retrieving helpful information from huge datasets. it's used in period of time applications like social site data process and medicine applications to handle massive volumes of knowledge sets typically large, sparse, incomplete, uncertain, complex or dynamic information set from multiple and autonomous sources. Massive data processing conjointly deals with storage structure of mined results, in order that the user will simply get the central plan and answers to the queries, from the mined results. Information slicing is performed to interrupt the association cross columns, however to preserve the association inside every column. Information slicing will be classified into static slicing, dynamic slicing, simultaneous dynamic slicing, quasi-static slicing and amorphous slicing. Clustering is additionally a basic task performed within the massive information mining process for the information discovery and getting patterns for use within the giant processing applications. This paper surveys about the large data processing techniques, information slicing techniques and clustering techniques and conjointly discusses concerning its blessings and drawbacks. The performance and quality activity of the big data mining

Dr P. Logeswari[1] ,Assistant Professor, Department of Computer Science[1],Sri Krishna Arts and Science College[1], Email: **logeswarip@skasc.ac.in**
J. GokulaPriya[2] , Research Scholar, Department of ComputeScience[2],Sri Krishna Arts and Science College[2], Email:**gokulapriyajaganathan1@gmail.com**
G.Banupriya[3] , Research Scholar, Department of ComputeScience[3],Sri Krishna Arts and Science College[3], Email: banu.snmv7@gmail.com
S.Sudha[4], Research Scholar, Department of ComputeScience[4]**,** Sri Krishna Arts and Science College[4], Email: sudhasw89@gmail.com
S.Sharmila[5], Assistant Professor, Department of Computer Science, Nallamuthu Gounder Mahalingam College, Email: mcasharmi2007@gmail.com

algorithms, mining platforms, and information slicing techniques and clump techniques are mentioned.

Keywords: Big Data Mining, Cloud Computing Technique, Clustering Techniques, Collaborative filtering, Data Mining and Data Slicing.

1. INTRODUCTION

Recent years have witnessed a dramatic increase in our ability to gather information from varied sensors, devices and freelance or connected applications, in several formats. This information flood has outpaced our capability to method, analyse, store and perceive these datasets. There occur vital challenges in leverage the large quantity of knowledge. {the information|theinfo|the information} mining method is meant to go looking consistent patterns or systematic relationships between the variables and validate the findings by applying the detected patterns to the new data set, for exploration of massive information. This is often extremely a difficult task because of the information quality and increase within the computation time.

Big information may be a presently rising example that promotes the infrastructure growth and development of connected information centre computer code. Huge information may be a computation-oriented method that stresses the storage capability of the cloud system. The key objective of the cloud computing technique is to supply huge information applications with fine-grained computing capability, by exploitation immense information computing and storage resources underneath intense management. Consequently, the emergence {of huge|ofmassive|of huge} information conjointly accelerates the event of cloud computing that has complete solutions for storing and process big information.

Effective management of massive information is achieved by the distributed storage technology. Huge information is no inheritable and analysed with efficiency, because of the parallel computing capability of cloud computing. With the fast development within the networking, information storage and information assortment capabilities, the appliance of massive information is currently chop-chop increasing in varied domains. Huge data processing permits retrieval of helpful information from this massive datasets. Usually, data processing is that the technique of analysing information from totally different prospects and

summarizing these information into fascinating, perceivable and helpful models. For higher deciding, the big repositories information of information} collected from totally different resources need a correct mechanism for extracting knowledge from the databases. Therefore an outsized decide to exploit these immense data processing architectures was initiated.

With the present technological advancement, the massive information and cloud computing technologies are positively and increasingly reticular with one another. Huge information operates within the higher level supported by the cloud computing and provides the functions kind of like those of info and economical processing capability. Ascent of application demand and cloud computing developed from the virtualized technologies have stirred up the evolution of massive information. Therefore, cloud computing not solely provides huge information computation nd process capability, however conjointly acts as a service mode. The advances of cloud computing promote the event of massive information, specified each technologies supplement one another. the most important reasons for the employment of the cloud computing technology within the huge information technology implementation ar low hardware price, low process price and massive information testing ability.

Security and loss of management ar the most important issues relating to the cloud computing technology. Big data processing conjointly deals with storage structure of well-mixed results, in order that the user will simply get the central plan and answers to the queries, from the well-mixed results. Information slicing is performed to interrupt the association cross columns, however to preserve the association inside every column. Information slicing is classified into static slicing, dynamic slicing, coincidental dynamic slicing, quasi-static slicing and amorphous slicing.

Clump is additionally a basic task performed within the huge {data mining data meth ding} process for the information discovery and getting patterns to be used within the massive processing applications. The clump technique involves hierarchal clump, partition clump, density based mostly and grid-based clump. This paper surveys regarding the massive data processing techniques, information slicing techniques and clump techniques. This survey discusses regarding the benefits and disadvantages of the massive data processing techniques, information slicing techniques and clump techniques.

The paper is organized as follows: Section II illustrates the massive data processing platforms, huge data processing algorithms, information slicing technique and clump technique. Section III describes the results and discussions and Section IV presents the conclusion of the survey.

2. BIG DATA MININGT ECHNIQUES

This paper surveys regarding the large data processing techniques, knowledge slicing techniques and bunch techniques. This survey discusses regarding the benefits and disadvantages of the large data processing techniques, knowledge slicing techniques and bunch techniques.

A. huge data processing Platforms:

1) MapReduce

MapReduce [1] could be a distributed Programming Model supposed for giant cluster of systems which will add parallel on an oversized dataset. The work hunter is to blame for handling the Map and cut back method. The MapReduce framework [2] sorts the outputs of the maps, that area unit then given as input to the cut back tasks. Each the input and output of the work area unit keep within the classification system. Because of parallel computing nature of MapReduce, parallelizing data processing algorithms victimisation the MapReduce model has received vital attention from the analysis community since the introduction of the model by Google. It's extremely onerous to write down the queries in Java or Python. The cut back section cannot begin till the completion of map section. This leads to the severe performance degradation.

2)Hbase

HBase[3] could be a column-based management system that's most typically used for large knowledge applications. In fact, the HBase [4] permits grouping of the many attributes into column families and storing all parts of the column family along. Hbase could be a distributed and climbable information for random browse or write operations and hosting of enormous tables. It additionally permits fast access and fault-tolerant storage of big quantity of distributed knowledge. However, storage of enormous size of binary files in HBase is incredibly tough. HBase is dear in terms of hardware needs and memory block

allocations.

3) Dynamo Amazon

DynamoDB [5] could be a fully managed information that supports each document and key-value knowledge models. Generator [6] creates an information table for storing and retrieving any quantity of knowledge and aiding many levels of request traffic. {the knowledge|theinfo|the information} and traffic for the information table will be mechanically contact adequate range of servers to tackle the requested capability level specified by the client and data storage capability, whereas maintaining the consistency and quick performance rate. However it supports solely a restricted set of queries and it's tougher to ascertain knowledge consistency. With Dynamo DB, the categorisation field's area unit to be set before the creation of the information, and it cannot be changed.

4) Asterix

Asterix [7] aims to mix the reliable principles from parallel information systems with those of the Web-scale computing community, like fault tolerance for long running jobs. Asterix [8] is planned to be a parallel and semi structured info management system with the aptitude for ingesting, storing, indexing, querying, analysing, and publication Brobdingnagian quantity of semi structured knowledge. Asteroid is well-suited to handle the rigid still as versatile and extremely complicated knowledge. It supports each system-managed dataset and external dataset. the most limitation of the Asteroid is that the absence of cost-based question optimizer.

5) Hadoop

Hadoop [9] is at the start perceived supported MapReduce, as a programing language supported Java. Within the Hadoop platform [10], the applying is softened into many tiny segments. The Hadoop writes applications that perform fast multiprocessing of big quantity of knowledge on giant clusters and it works on MapReduce programming model that could be a generic execution engine that parallelizes computation over an oversized cluster of machines. Hadoop will give abundant required strength and measurability choice to a distributed system as Hadoop provides cheap and reliable storage. However, Hadoop isn't the simplest answer for the

organizations that touch upon touch of knowledge. There are cascading failure issue and multi-tenancy issue within the Hadoop platform.

B. massive data processing Algorithms:

Big data processing algorithms [11] area unit wont to discover data or extract pattern from {the large|themassive|the massive} information set or big information set. Numerous algorithms used for large data processing area unit Two-section top-down Specialization (TPTDS) approach, Tree- primarily based Association Rules (TARs), Fuzzy C-Means (FCM) algorithmic program and Associate Rule Mining (ARM) algorithmic program.

1) Two-Phase top-down Specialization (TPTDS) approach

The TPTDS approach [12] is especially used for mining great amount of knowledge with high privacy. The 2 phases of this approach area unit job level and task level. Parallelization is achieved in each the phases. The essential plan of TPTDS [13] is to achieve high quantifiability by creating a exchange between quantifiability and information utilit y. Despite its well usage for the privacy preservation on privacy sensitive giant scale information sets, the most limitation of the TPTDS approach is that it cannot give privacy preservation for the info set with giant quantifiability.

2) Tree-Based Association Rule (TAR)

Tree-Based Association Rule [14] is employed for mining the extensile terminology (XML) documents and therefore the obtained results are often keep in XML formats. Association rules [15] describe the co-occurrence {of information of knowledge|of information} things in a very great amount of collected data. The standard of Associate in Nursing association rule is measured by means that of support and confidence. The association rule is extended within the context of relative databases to create it adapt for ranked nature of XML documents. Relationships among sub trees of XML documents are often found with the matter contents of leaf components and attributes price. Updatability of the document storing TAR throughout the amendment happens in original XML information sets and also as their index may be a limitation for Tree-Based Association Rule.

3) Fuzzy C-means (FCM)

FCM algorithmic program [16] is employed for cluster of terribly giant information for the method of mining from labelled information, giant image information and unloadable information. The cluster approaches will accomplish 2 objectives like acceleration for loadable information and approximation for unloadable information. Fuzzy partitions area unit additional versatile than the crisp partitions therein every object will have membership in additional than one cluster. FCM [17] is initialized by selecting the objects indiscriminately from the dataset to function the initial cluster centres, the algorithmic program terminates once there area unit solely negligible changes in cluster centre locations. Fuzzy C-Means algorithmic program has many limitations as developing and investigation ascendible solutions for terribly giant

fuzzy cluster. To look at wherever kernel solutions are often used, it's potential to use cluster validity indices to settle on the suitable kernel. Quality of cluster is measured by requiring full access to the objects vector information.

4) Associate Rule Mining (ARM)

ARM [18] may be a technique for revealing meaningful relations between variables in information bases, for clinical data processing applications. ARM [19] is wide employed in applications like heart condition prediction, aid auditing and medical specialty diagnosing in hospitals. 2 necessary metrics area unit support and confidence that assess the frequency and level of association of a rule. So as to find frequent and assured association rules, the mining method needs the users to specify minimum support and minimum confidence values as thresholds. The most goal of ARM is to search out all frequent and assured rules supported these 2 users' nominative values. The most limitation of the ARM is manual specification of variables for interest and cut points by clinicians. Table.1 shows the comparison of the massive data processing algorithms.

TABLE.1 COMPARISON OF BIG DATA MINING ALGORITHMS

Algorithm/ Approach	Performance Criteria	Usage
Two-Phase Top-Down Specialization (TPTDS) approach	Execution time and Information Loss	Privacy Preservation Of data
Tree-Based Association Rules (TAR) approach	Extraction time and Answer time	Mining from semi structured (XML) document
FCM Algorithm	Run time	Clustering of data
Associate Rule Mining (ARM)	Comorbidity	Mining from ICU(clinical) data

C. information Slicing Technique

Data slicing [20] performs breaking of the association cross columns, however to preserve the association inside every column. Slicing performs each horizontal and vertical partitioning of the dataset. This reduces the info spatiality, whereas conserving the higher information utility level. Information slicing includes vertical partitioning and horizontal partitioning of dataset. Vertical partitioning is finished by grouping the attributes into columns supported the correlations among the attributes. Every column contains a set of attributes that square measure extremely correlative. Horizontal partitioning is finished by grouping tuples into buckets. At last, inside every bucket, values in every column square measure haphazardly permutated to interrupt the association between completely different columns. Information slicing will be classified into static slicing, dynamic slicing, coinciding dynamic slicing, quasi-static slicing and amorphous slicing.

1) Static Slicing

Static slicing [21] is performed while not creating assumptions concerning the input of the program. Dependency among completely different field parts for unified modeling language (UML) model is known for computing static slices. The slices[22] square measure computed by gathering statements and management predicates by method of a backward traversal of the program's management flow graph (CFG) or program dependence graph(PDG), beginning at the slicing criterion. However, a static slice could contain statements that haven't any influence on the values of the variables of interest for the actual execution. Throughout execution of a

program, the worth inputted could cause surprising result.

2) Dynamic Slicing

Dynamic slicing [23] takes the input equipped to the program throughout execution and therefore the slice contains solely the statement that caused the failure throughout the particular execution of interest. Dynamic slicing [24] uses dynamic analysis to spot all and solely the statements that have an effect on the variables of interest on the actual abnormal execution trace. Dynamic slicing can treat every component of Associate in nursing array on an individual basis, whereas static slicing considers every definition or use of any array component as a definition or use of the whole array. However, dynamic slice is additionally immense for a sensible massive software package. The dynamic slicing technique produces viable slices that square measure correct on just one input.

3) Simultaneous Dynamic Slicing

Simultaneous dynamic slicing [25] on a group of take a look at cases isn't merely given by the union of the dynamic slices on the element take a look at cases. Indeed, merely the union of dynamic slices is unsound, therein the union will not maintain coinciding correctness on all the inputs. Associate in nursing repetitious formula is given that, beginning from Associate in nursing initial set of statements, incrementally construct the coinciding dynamic slice, by computing the iteration a bigger dynamic slice. This approach will be employed in program comprehension for the isolation of the set of the statements like specific program behavior. Coinciding dynamic slicing will be thought-about as a refinement of ways for localizations of functions supported take a look at cases, as a result of it takes into consideration the info flow of the program then permits the reduction of the set of chosen statements.

4) Quasi-Static Slicing

Quasi-static slicing [26]is a hybrid of static and dynamic slicing. In Quasi-static slicing [27], price|the worth} of some variables square measure mounted and therefore the program is analyzed throughout variation within the value of alternative variables. The behavior of the first program isn't modified with relation to the slicing criterion. Slicing criteria includes the set of variables of interest and initial conditions and thus similar slicing is named as Conditioned slicing. this can be Associate in Nursing economical methodology for program comprehension.

But, this system fails to prove uniform treatment of static and dynamic slicing.

5) Amorphous Slicing

Amorphous slicing [28] depends on protective the linguistics of the program, like the slicing criterion. The slices fashioned don't seem to be as massive because the alternative slicing techniques. The slice is significantly simplified sort of the program with relation to the slicing criterion. Amorphous slicing assists in program comprehension, analysis and reprocess. However once amorphous slicing is applied on the ultimate worth of the variable biggest, the linguistics is preserved by following the constant folding technique to realize the slicing. The amorphous slicing isn't created by deleting impertinent statements however performs doable syntax transformation conserving the linguistics with relation to the slicing criterion. the result of this slice can ne'er be larger than the first program to be sliced.

D. bunch technique

Clustering technique [29] is the essential task of the info Mining method, for categorizing or grouping similar information things along to cut back the info quantity. Numerous algorithms square measure used for bunch. The foremost oft used bunch techniques square measure Partitioning-based, Hierarchical-based, Density-based and Grid-based ways.

1) Hierarchical cluster methodology

Hierarchical cluster [30] performs gradual construction of the clusters, by merging the smaller cluster into larger ones or rending the larger clusters. The cluster tree known as asdendrogram showing the connection of the clusters is created because the end product of the algorithmic rule. The information things square measure clustered into disjoint teams, by cutting the dendrogram at a desired level. Stratified cluster will be loosely classified into agglomerate stratified cluster and discordant stratified cluster. Within the agglomerate approach, every information point's square measure thought-about to be a separate cluster and therefore the clusters square measure unified on every iteration supported criteria. Within the discordant approach

all information points square measure thought-about as one cluster and splatted into range of clusters supported sure criteria. The most disadvantage of the stratified cluster methodology is unclearness of termination criteria, inability to create corrections throughout the splitting/merging method and lack of interpretability concerning the cluster descriptors. Severe effectiveness degradation in high dimensional areas happens because of the curse of spatial property development.

2) Partitioning cluster methodology

Partitioning cluster methodology [31] performs direct decomposition of information set into a group of disjoint clusters. The most downside in cluster is that the vital alternative of the quantity of clusters and rising of various forms of clusters. Formatting of the centroids of the cluster might also be crucial. Some clusters could also be left empty if their centroids square measure placed distant from the information distribution. Hence, to beat these limitations, the partitioning technique divides information objects into variety of partitions representing a cluster. Severe effectiveness degradation in high dimensional areas as the majority pairs of points is regarding as distant as average. The construct of distance between points in high dimensional areas is ill-defined. The partitioning cluster methodology is very wise to formatting section, noise and outliers and enable to alter non-convex clusters of variable size and density.

3) Density-based cluster methodology

The density-based cluster technique [32] separates the information objects, supported the density regions, property and boundary. Density based mostly algorithmic rule outlined as a connected dense element still grow within the given cluster as long because the density within the neighborhood exceeds a definite intensity level. Therefore, the density-based algorithms will discover the capricious form clusters and supply protection against outliers and noise. Therefore the density purpose is calculated for determinative the dataset functions that influence a selected information. The most disadvantage of this approach is its high sensitivity to the input parameter setting and poor cluster descriptors.

4) Grid-based cluster methodology

The grid-based cluster approach [33] performs division of object house into a finite range of cells that type a grid structure to perform the cluster operations. This approach

depends solely on the quantity of cells in every dimension within the amount house and doesn't rely upon the quantity of information objects. This grid based mostly approach goes through the dataset to calculate the applied math worth for the grids. Therefore the interval of the grid based mostly cluster approach is fast. The performance of the grid-based methodology is improved, since it depends on the dimensions of the grid that is way smaller than the dimensions of the info. However, usage of single uniformgrid isn't decent to get the specified cluster quality for extremely irregular information distributions.

3. RESULTS AND DISCUSSIONS

Various techniques for large {data mining|data meth ding} process square measure illustrated during this section. The performance and quality measuring of the massive data processing algorithms, mining platforms, information slicing techniques and cluster techniques square measure mentioned during this section.

TABLE 2. INFORMATION ABOUT VARIOUS TECHNIQUES INVOLVED IN BIG DATA MINING PROCESS.

Techniques	Author& Reference	Year	Performance	Quality Measurement
Big Data Mining Platforms				
Map Reduce	J.Li et al [1]	2012	The key benefit of MapReduce is that it automatically handles failures, hiding the complexity of fault-tolerance from the programmer.	1. Co-schedulingspeedup 2. Executiontime 3. Number ofprocessors
	H.Wang et al [2]	2012	This new approach improves the ability to scale up the Hadoop clusters to a much larger configuration. In addition to this, it also permits parallel execution of a range of programming models.	1. Feature extractiontime 2. Clusteringtime 3. Averageaccuracy 1. Mean averageprecision
Hbase	Aksu et al [3]	2013	Hbase supports random and fast insert, update and delete access.	1. K-core constructiontime 2. Executiontime 3. Maintenanceoverhead
	Rabl et al [4]	2012	Hbase provides linear and modular scalability, strictly consistent data access, automatic and configurable sharing of data.	1. Throughput 2. Read latency 3. Writelatency

Dynamo	R.Gupta et al [5]	2012	Dynamo provides incrementalscalability. Hence, keys are partitioned dynamically using a hash function to distribute the data over a set of machines ornodes.	-
	Moharil et al [6]	2014	Dynamo provides faster and predictable performance with seamless scalability with minimal database administration.	1. Number of executors and tasks 2. Processingtime 3. Linestailed 4. Number ofclusters formed
Asterix	Behm et al [7]	2011	Asterix supports large, self-managing data sets and index structures as well as query planning, processing, and scheduling approaches that are scalable and adaptable to highly dynamic resource environments.	1. Speedup Ratio 2. Number of nodes in cluster 3. Trainingsize 4. Loss on testdata
	Kaldewey et al [8]	2012	Since ASTERIX provides rich spatial support, spatial aggregation queries are executed efficiently by using a secondary R-tree index. Thus, all records outside of the query bounding region are filtered quickly.	1. Hadoop distributed file system (HDFS) read bandwidth 2. Diskbandwidth
Hadoop	Dede et al [9]	2013	The design of Hadoop achieves data locality and considerable performance improvement by placing data on the compute nodes.	1. Check pointinterval 2. Overhead 3. Number oftasks 4. Number of inputrecords 5. Processingtime
	Zhang et al [10]	2013	Hadoop can effectively reduce the search time and improve the retrieval speed.	2. DataCapacity 3. Cluster Processingtime 4. Single node costtime 4. Cluster costtime
Big Data Mining Algorithms				
Two-Phase Top-Down Specialization (TPTDS) approach	X.Zhang et al [12]	2014	The two-phase TDS approach gains high scalability via allowing specializationsto be conducted on multiple datapartitions in parallel.	1. Executiontime 2. Informationloss
	X.Zhang et al [13]	2014	It gains high scalability and efficiency of sub-tree anonymization scheme to anonmize data sets for privacy preservation.	1. Privacy preservingcost 2. Privacy leakagedegree 3. Number of datasets
Tree-Based Association Rule (TAR)	A. R. Islamand T.S.Chung[14]	2011	Tree–Based Association Rules performs mining all frequent association rules without imposing any prior restrictionon the structure and the content of therules.	1. Extractiontime 2. Answer Time

	K.S. Rani et al [15]	2013	Enables the user to extract efficient answering from the XML documents.	1. Number of clusternodes 2. Incremental eventdata size 3. Executiontime
Fuzzy C-means(FCM)	S.Kannan et al [16]	2012	The Fuzzy C-Means algorithm supports the clustering of very large data or big data.	1. Runtime
	H. Izakian and A. Abraham[17]	2011	The Fuzzy C-means algorithm is efficient, straightforward and easy to implement.	1. Number of clusters
Associate Rule Mining (ARM)	F. Suchanek and G. Weikum[18]	2013	Associate Rule Mining algorithm generates quantitative and real time decision support rules for Intensive Care Unit (ICU) by predicting the characteristics of ICU stay.	1. Comorbidity
	Galárraga et al [19]	2013	The associate rule mining algorithm achieves improved run time, quality of the output rules and reasonably predicts the precision of the rules.	1. Aggregatedpredictions 2. Aggregated precision
Data Slicing				
Static Slicing	Alomari et al [21]	2012	The approach is highly scalable and can generate the slices for all variables of the Linux kernel in less than 13 minutes.	1. Slicesize 2. Systemsize
	Santelices et al [22]	2013	The accuracy of slices is improved. Slicing can be done within a short period.	1. Percentage of slice inspected 2. Percentage ofimpacts found 3. Slicingranking 4. Run time overhead
Dynamic Slicing	J. T. Lallchandani and R. Mall[23]	2011	The advantage of dynamic slicing isthe run-time handling of arrays and pointer variables.	1. Reverse executiontime 2. Selection sortrange 3. Averagespeedup 4. Average code size reduction 5. Memoryoverhead
	J. Zhong and B. He[24]	2014	Dynamic slicing produces a more compact and precise slice. Reverse execution along a dynamic slice skips recovery of unnecessary program state.	1. Sliced executiontime 2. Slicesize 3. Workload 4. Probability

Simultaneous Dynamic Slicing	M. A. El-Zawawy[25]	2014	As the slice size is dynamically calculated, the lines of code do not affect the slice criteria. Hence, simultaneous dynamic slicing achieves better performance under all the conditions. The time required for calculating the slice size is reduced, since the slice size is calculated on the runtime.	1. Lines ofcode 2. Slicesize 3. Distance 4. Percentageaffect 5. Functionpoints
Quasi-Static Slicing	Swain et al [26]	2012	It allows a better decomposition of the program giving the maintainer the possibility to analyze code fragments with respect to different perspectives.	1. Number of testcases 2. Slice testcoverage
	S. Koushik and R. Selvarani[27]	2012	Quasi-static slicing allows a better decomposition of the program giving human readers the possibility to analyze code fragments with respect to different perspectives.	1. Executiontime 2. Performancegain 3. Bounded model checking (BMC)
Amorphous Slicing	Androutsopoulo et al [28]	2013	It produces slice of smaller size.	1. Slicesize 2. Number of lines of codes forslice 3. Executiontime
Clustering Technique				
Hierarchical Clustering Method	M.Verma et al [30]	2012	Hierarchical clustering method is more versatile and easy to handle any forms of similarity or distance. It is consequently applicable to any attribute types.	1. Number ofclusters 2. Cluster instances
Partitioning clustering Method	K. Aparna and M. K. Nair[31]	2015	The partition-based technique utilizes an iterative way to create the cluster.	1. Number ofclusters 2. Datapoints
Density-based Clustering Method	Kriegel et al [32]	2011	This approach enables discovery of arbitrary-shaped clusters with varying size. It is more resistance to noise and outliers.	1. Number ofclusters 2. Cluster instances 3. Number ofiterations 4. Time taken to buildmodel 5. Loglikelihood
Grid-based Clustering Method	E. G. Mansoori[33]	2014	The grid-based clustering technique requires low processing time, since it depends on the size of the grid instead of	1. Timeconsumption 2. Clustering correctrate 3. Noise filteringrate
			the size of the data.	

4. CONCLUSION

Big data processing conjointly deals with storage structure of well-mixed results, so the user will simply get the central plan and answers to the queries, from the well-mixed results. Knowledge slicing is performed to interrupt the association cross columns, however to preserve the association among every column. cluster is additionally a basic task performed within the massive data {processing} process for the information discovery and getting patterns to be used within the massive processing applications. The performance and quality activity of the massive data processing algorithms, mining platforms, knowledge slicing techniques and cluster techniques area unit mentioned.

References

1. J. Li, P. Roy, S. U. Khan, L. Wang, and Y. Bai, "Data mining using clouds: An experimental implementation of apriori over mapreduce," in *12th International Conference on Scalable Computing and Communications (ScalCom)*,2012.
2. H. Wang, Y. Shen, L. Wang, K. Zhufeng, W. Wang, and C. Cheng, "Large-scale multimedia data mining using MapReduce framework," in *CloudCom*, 2012, pp.287-292.
3. H. Aksu, M. Canim, Y.-C. Chang, I. Korpeoglu, and O. Ulusoy, "Multi-resolution Social Network Community Identification and Maintenance on Big Data Platform," in *IEEE International Congress on Big Data (BigData Congress),* pp. 102-109,*2013.*
4. T. Rabl, S. Gómez-Villamor, M. Sadoghi, V. Muntés-Mulero, H.-A. Jacobsen, and S. Mankovskii, "Solving big data challenges for enterprise application performance management," *Proceedings of the VLDB Endowment,* vol. 5, pp. 1724-1735,2012.
5. Dr. P.Logeswari "Extraction of Subset- Want in Data Stream using EMDMICA Algorithm " Volume 7 Issue VI, June 2019.
6. Dr. P.Logeswari, J.Gokulapriya "A Literature Review on Data Mining Techniques "in July Volume -7 Issue -7.
7. .Dr. P.Logeswari, J.Gokulapriya "Literature Survey on Big Data mining And Its Algorithmic Techniques "in July Volume -8 Issue7.
8. Dr. P.Logeswari, G.Banupriya "A Survey on Implementations Solutions for Attack Prevention Cryptography Technique's in WSN UsingNS2" Volume 7,Issue 6 June 2021.
9. Dr. P.Logeswari, G.Banupriya "Review on Cryptography Techniques in WSN for Attack Prevention" volume 8, Issue 8.
10. Dr. P.Logeswari, J.Gokulapriya "Data Mining Approaches and Applications" (Conference Paper).

11. Dr. P.Logeswari, G.Banupriya “Image processing and its Application” (Conference Paper).
12. Dr. P.Logeswari, G.Banupriya “Cryptography Techniques and its analysis” (Conference Paper).
13. Dr. P.Logeswari, S.Sudha “Survey on Privacy Preserving Secure Mining” (Conference Paper).
14. Dr. P.Logeswari, J.Gokulapriya “Analysis of Data Mining Techniques and its Application” (Conference Paper).
15. Dr. P.Logeswari, G.Banupriya “Cryptography Techniques and its Analysis” (Conference Paper).
16. Dr. P.Logeswari, S.Sudha “A Survey on Privacy Preserving in Data Mining”Volume-7, Issue-8 August 2021.
17. Dr. P.Logeswari, S.Sudha “A Review on Privacy Preserving in Data Mining” Volume-8, Issue-6 June2021.
18. Dr. P.Logeswari, J.Gokulapriya “Research Paper on Big Data and Hadoop” (Conference Paper).
19. Dr. P.Logeswari, J.Gokulapriya “Data Mining for Security Applications” (Conference Paper).