ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 12, No. 2, 2017, *pp*.141-150

Adaptive Cluster Multi Dimensional Data Analysis in Map Reduce Framework using Matlab

Uma Mahesh Kumar Gandham[1], Dr P Suresh Varma[2]
[1] Research Schalor, AKNU University, Asst. Professor, Dept of CSE, GIET Engg. College, Rajamahendravaram – AP, India, 533296, E-mail: umamahesh.gnadam@gmail.com.
[2] Professor, Dept of CSE, University College of Engineering, Adikavi Nannaya University, Rajamahendravaram – AP, 533296, India.
(Received January 6, 2017, accepted March 07, 2017)

Abstract. Data privacy protection is one of the most disturbed issues on the present industry Data isolation issue require to be addressed immediately previous to the data sets are common on a cloud. Data point refers to as hiding compound data for owner of data records. Expand the process of analysis over big multidimensional information as well, by importance open problems and real investigate trends. In this research new algorithm called Adaptive Cluster Multi Dimensional Data Analysis in Map Reduce Framework is been implemented on mat lab. Dispensation great quantity of information is attractive a confronting for data investigation software. Data clustering is a documented information study technique in data mining while adaptive K-Means is the well known partition clustering method. The inspiration at the back ACMDDA proposing algorithm is to contract with different dimensional information clustering, with minimum amount error rate and utmost meeting rate. With the help of multidimensional data set of map dipping framework, here implemented algorithm will amplify the competence of the big data for out system.

Keywords: Multi Dimension; Cluster; Data analysis; Matlab; K-means; Map Reduce.

1. Introduction

Penetrating for helpful data flow of in sequence in the middle of huge amount of data is known as the field of big data. Data is characteristically broken down for the study for a data warehouse into different dimensions such as instance period, invention segment and the environmental location. Capacity is estranged into classes a lot of statistics are currently being utilize[1][2], handle and dissect as new, automatic and much different structure since of a boom in the ground of computerization and digitization events. Every time explanation we need with a low-cost storage space that will allow vast commerce dangerous application and data. This paper is part of this multidimensional analysis which specially works on preserve the seclusion of rising individual information. To avoid the exposé of personal information's unique individual identifiers like individual numbers, social safety figure or any other single information can easily be deleting from datasets previous to releasing them widely [3].

Huge compute cluster are more and more individual used for data investigation. The data level and cost of this cluster make it serious to get better their in service competence, counting energy [4]. This paper focuses on an exchange use case what we call Map Reduce with Interactive Analysis (CMIA) Cluster Map Reduce with Interactive analysis workloads. CMIA workloads hold interactive services, customary lot dispensation, and large-scale [5]. latency-sensitive dispensation. CMIA workloads need a very dissimilar move toward to energy-efficiency, one that focuses on lessening the quantity of power used to repair the workload. "Big Data" refers to huge amount of formless data shaped by high-performance application lessening in a wide and varied family of request scenario: from technical compute application to social networks, from e-government application to health check in order systems.[6][7][8]

Yanpei Chen et.al [4] proposed to such workloads run on big clusters, size and cost make energy. Mechanism on Map Reduce power efficiency has not yet careful this workload class. Increasing hardware operation helps get better competence, but is demanding to attain for MIA workloads. BEEMR achieves 40-50% force investments below tight plan constraint, and represent a first step towards improving energy competence for a more and more significant class of datacenter workloads. **Alfredo et.al [3]** provides an impression of state-of-the-art data selection issue and achievement in the field of analytics over big data, and we make bigger the conversation to analytics over big multidimensional data as well, by stress open problems and real investigate trends. This plays a most important position in next-generation Data

Published by World Academic Press, World Academic Union

DOI https://doi.org/2024-JICS-22490 | Generated on 2025-04-20 19:32:59
OPEN ACCESS

Warehousing and OLAP research. **Pramod Patil** et. al[7] proposed a big data anywhere information is raising twice by its size over year and year. So it is very hard to knob and procedure the huge quantity of data. Data storage space and data treatment be supposed to be done in real time and without loss of data. Cloud compute resolve the complexity of storage space and ease of use for data investigation task.

Map Reduce has be seen as one of the key enable approach for gathering incessantly rising stress on compute capital compulsory by huge data sets only on one dimensional issue so this huge difficulty at present situation. The cause for this is the far above the ground scalability of the Map Reduce example which allows for especially similar and dispersed implementation over a large number of compute nodes. Big data and parallel compute frameworks comes into picture where data investigation work need to be approved out. Here we make obvious that; unlike obtainable technique we can able to examine the millions of tulles in real time for our datasets.[9]10]

2. Proposed methdology

Multidimensional data analysis application is mainly used in social network system. Enterprises produce massive quantity of data every day. These data are pending from a variety of aspect of their products, for example the overall data. The raw data is extract, distorted, cleanse and then stored under multi-dimensional data model, such as star-schema1or hybrid prototype. Those queries are usually complex and involve large-scale data access. Here are some features are summarizing for multidimensional data analysis queries as below:[6]9]

- Dimensional adaptive embedding technique presents a cluster subspace of multidimensional data space in a hierarchical cluster mechanism.
- Multi view dimensional technique with map reduce to attributes to coordinates such as scatter plot matrix, web based hyper slide and hyper box.[7]
- Intelligent dimensional reduction technique map reduce space of lower dimension with preserving relationship of multidimensional data set.[11]

2.1. Work flow

Data mining technique create potential to examine and determine information of data sets. However, the custom clustered data are not as long as more correct data for large datasets. Map reduces hold up for implement cluster algorithms by conduct large quantity of data in addition with Matlab framework.[5] Using Map Reduce encoding replica for dispensation the data cluster in dispersed system. To get better the presentation of the large-scale datasets clustering on the on its own computer. To find the correctness of data in Intelligent K- Mean's algorithm to work out MSSE (Mean Sum Square error) charge base upon Euclidean detachment using Map Reduce framework for 2dimension and 3 dimension datasets [8]. The work flow diagram of the proposed method is shown in figure 1.

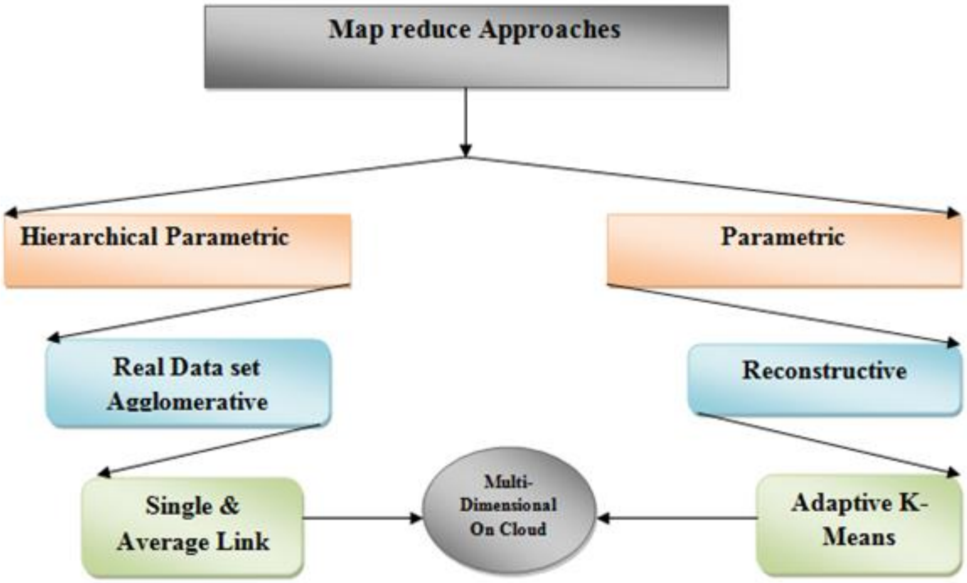


Fig.1. Work flow diagram of the proposed method

DOI https://doi.org/2024-JICS-22490 | Generated on 2025-04-20 19:32:59

OPEN ACCESS

2.2. Proposed method features

Big data computing Average to speed up the process in data mining concept over map reduce. Stability analysis in over all multi dimensional data set pre-determined Map reduces. Check the advance properties of Adaptive K-Means Cluster in different steps to be executed. Finally check speed and memory of overall execution [12]. The database process consists of

2.2.1. Data cleaning

It is from time to time also known as data cleansing. It is a pace in which noisy or immaterial data are detached from the compilation of the database. Fill in absent values, flat noisy data, recognize or take away outliers, and decide inconsistency addition of manifold databases, data cubes, or files Normalization and aggregation obtain abridged symbol in quantity but produce the similar logical consequences part of data decrease but with exacting significance, particularly for arithmetical data Use the quality mean to fill in the absent value, or use the quality denote for all sample belong to the same class to fill in the missing value.[13][14]

2.2.2. Multi Dimensional Pattern evaluation

Pattern judgment is the majority computationally classy step in the procedure of fault detection. In this research, plan a pattern finding algorithm based on Google Map Reduce on cloud to get better the competence. Presentation assessment shows our algorithm can facilitate the discovery of larger dataset in large size network and has good scalability Advanced Map Reduce-based pattern finding (AMRPF) structure aim to put into practice recurrent pattern finding on intricate graphs based on Multi dimensional MatLab. Though it also works well on undirected graphs, here we still focus on introduce its request on heading for graphs. It’s more attractive and envoy to be relevant this structure on heading for graphs. For clearly depict AMRPF [15][16].

2.3. Cluster based Map Reduce programming

In Map Reduce framework process has two divide steps Map and Reduce steps. Each step is procedure on sets of (key, value) pairs. Whereas, the time of agenda implementation is alienated into a Map and a Reduce stage, every alienated by data move between nodes in the K –means cluster. In Mapped reason can decide the data main beliefs as contribution apply the reason to each worth to the known datasets and generate an output set.[17]

The map reduces in the form of (key, value) pairs. The framework then sorts the mapped purpose output and inputs them into a Reducer. This data move between the Mappers and the Reducer. The standards are communal at the node organization for that key. In algorithm map reduce phase produce another set of (key, value) pairs as final output. [7]

Adaptive K-means Clustering is a procedure of group with alike objects. Any cluster is supposed to show two main properties that fit in to, low inter-class resemblance and high intra class resemblance. This type Clustering technique used to collection a big figure of belongings jointly into clusters that split some resemblance. It’s a method to find out pecking order and arrange in a big or hard to appreciate datasets and in that way reveal are attractive pattern or make the data set easier to understand. Cluster investigation is used in many numbers in frame work over cloud.[15][18]. The working mechanism of the proposed approach is shown in figure 2.

DOI https://doi.org/2024-JICS-22490 | Generated on 2025-04-20 19:32:59

OPEN ACCESS

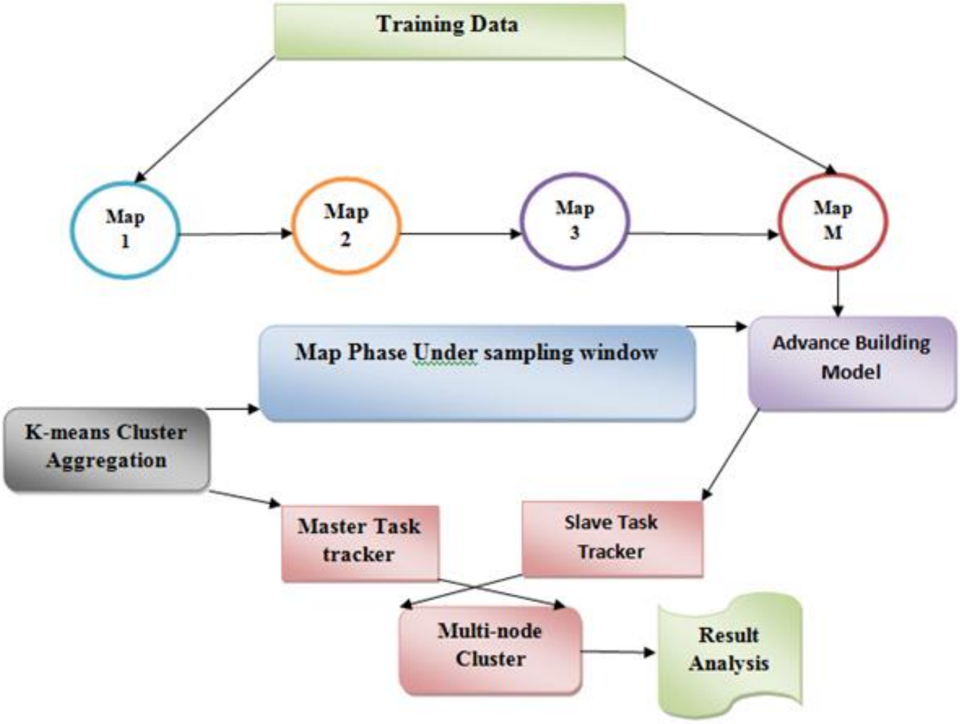


Fig.2. Proposed approach working Mechanism

The appraisal thus far illustrates that decrease slots are utilized less than map slots. unreliable map reduce ratio (i.e., raising the figure of map slots and lessening the number of decrease slots while custody cluster size steady) be supposed to allow map tasks in each lot to total faster without moving decrease tasks conclusion rates so leads to energy competence improvement, particularly for the latency-bound algorithm [19]20].

2.3.1. Competence issue of Map Reduce in Multi Dimensional

The good method of Map Reduce is a substance of contest for data investigation. The basic reason is that Map Reduce is not originally intended for data study scheme over prearranged data. The characteristic calculation anxious in Map Reduce is scan over a lot of shapeless data, like web page. In dissimilarity, the data study application characteristically access set in order. In the customary data analysis, the data-loading stage is to load data by a pre-defined schema. Still, the optimization, like directory and materialize view can indisputably pick up the presentation. Particularly, this optimization only need to be done one-time, and are reusable for all dispensation of query [21]. The flow chart of the proposed approach is shown in figure 3.

DOI https://doi.org/2024-JICS-22490 | Generated on 2025-04-20 19:32:59

OPEN ACCESS

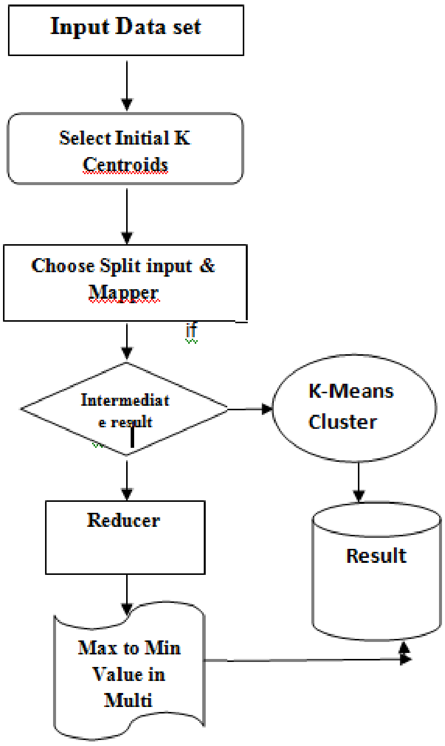

Input Data set
Select Initial K Centroids
Choose Split input & Mapper
if
Intermediat e result
K-Means Cluster
Reducer
Result
Max to Min Value in Multi

Fig.3. Flow chart of the proposed method

2.3.2 Proposed methodology algorithm

```
Step 1: Initial step to assign input dataset: d
Step 2: Select the group data to form G
Step 3: take Sampling ratio Sr
Step 4: Assign output cluster C
        // to take some model data in map reduce*****//
Step 5: RM mapper reduce read the data and send the element to one to another reducer with
        probability Pr
Step 6: to find cluster FC
Step 7: received elements in and passes cluster into M mappers
Step 8: R reducer use the elements from cluster from group G
begin
        Step 9: to merge the cluster received
        Step 10: return cluster C
end
```

2.3.3 Submissive Clustering in Map Reduce Framework analysis

K-means clustering method for information in five or more size, known as subspace cluster methods, more often than not follow one of two approaches density-based and k-means-based. A new review is found in Density-based method assume that a cluster is an information space region in which the constituent sharing is dense. Every area may have a chance form and the basics in it may be arbitrarily discrete. Map Reduce is an indoctrination structure to procedure big level information in a particularly similar way.[7][10] Map Reduce has two main compensation the programmer is unaware of the particulars connected to the data storage space, sharing, duplication, load balancing, etc and furthermore, it adopts the familiar concept of functional programming.[11]. The ACMDDA is listed below

Algorithm 1: ACMD2A Algorithm:

```
    Step 1: Input and Output Parameters I/O
    Step 2: Map Reducer r
    Step 3: Advance Mapper m
    Step 4: Adaptive Cluster C
        5:***** compute the more Dimensional data set in following cluster groups*****
    Step 5: analysis Map reduce frame work analysis C<C1<C2<C3= r1>r2>r3>r
```

DOI https://doi.org/2024-JICS-22490 | Generated on 2025-04-20 19:32:59

OPEN ACCESS

```
Step 6: so cluster parallel and equal with reduce to Map frame work.
Step 7: if C> r reducer then
Step 8: Cluster = result of S and sampling and ignore ration.
Step 9. Else cluster = result of Map reduce frame work.
Step 10: end if
Step 11: return cluster
```

2.4. K-means Cluster Design for Multi Dimensional Data Analysis using Matlab tool

Multidimensional information makes available add-on value to big data analytics. In this admiration, plan compound analytics over mat lab -included multidimensional data plays a serious role. Real analytics, though fairly well-developed, motionless do not go further than classical BI machinery, like diagram, plot, dashboard, and so forth, but multifaceted BI process of very great organization command for additional maybe by integrate main beliefs and consequences of different-in-nature discipline like figures. [22]

Revelation issue stand for a most important problem in Data Warehousing and OLAP research. This issue gets not as good in the background of big multidimensional data analytics, as here apparition must keep a stronger decision-support value. More multifaceted technique, such as multidimensional space examination approach, must be investigated to this end.[23][24]

3. Simulation and result analysis

Test for K-means clustering in mat lab used well know Machine learning analysis. A compilation of data base which is widely used by the researcher community of machine learning especially for the algorithm analysis of the discipline. With look upon to connections flanked by group that the algorithm makes in the test conduct, it was experiential that the majority important change occurs from first to second iteration. For all cases, in the first step all points are situated (100%), the third column include the figure of point to be exchange in the second iteration and the fourth column is the proportion dissimilarity in the figure of items exchange stuck between the first and second iteration.[6][15]. The projection of the mean proposed approach is shown in figure 4 and the projection of the cluster proposed approach is shown in figure 5.

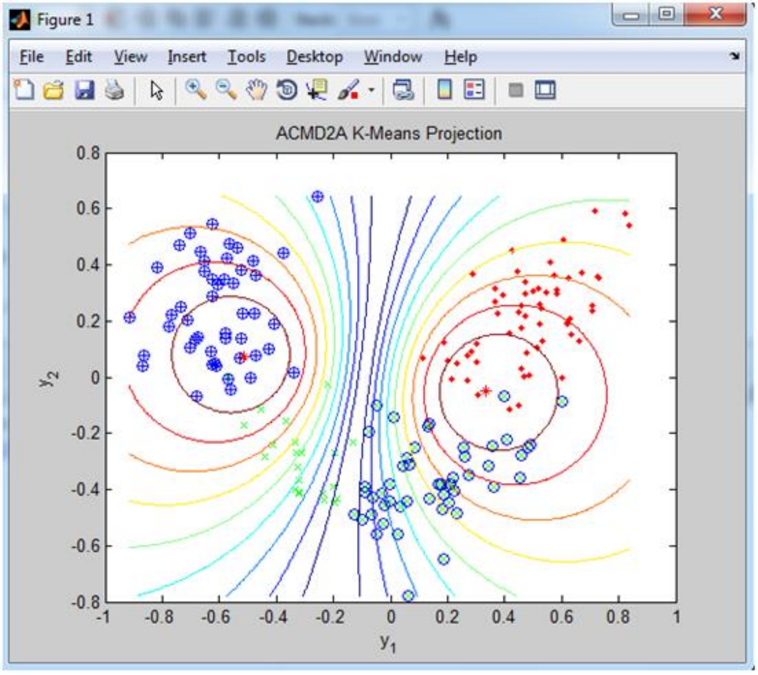


Fig.4. Mean base Map Reduce in ACMDDA Proposed Algorithm

DOI https://doi.org/2024-JICS-22490 | Generated on 2025-04-20 19:32:59

OPEN ACCESS

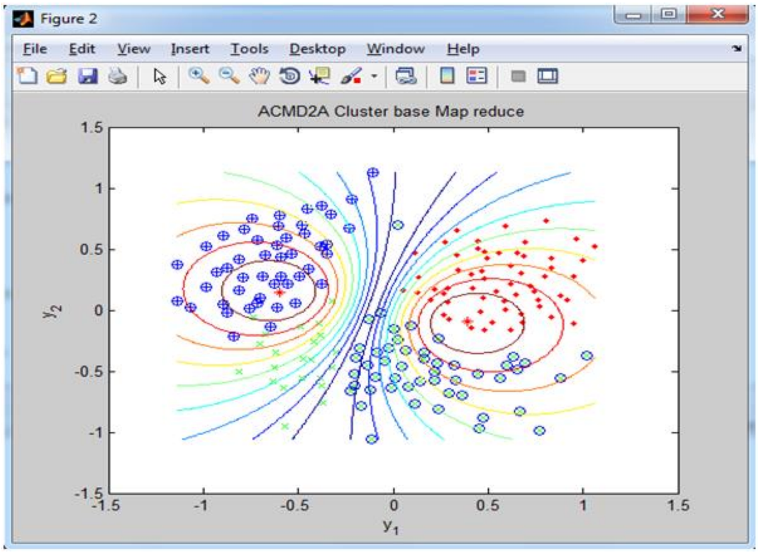


Fig.5. cluster base Map Reduce in ACMDDA

Cluster base map reduce which implies the status of different cluster group in sense of the overall prototype mechanism the cluster group had been separated in different color due to this overall huge data set map reduce is obtained very effectively. The axis Y1 and Y2 which represent the Dimensional and Data Set points. The multi dimension approach generated result is shown in figure 6.

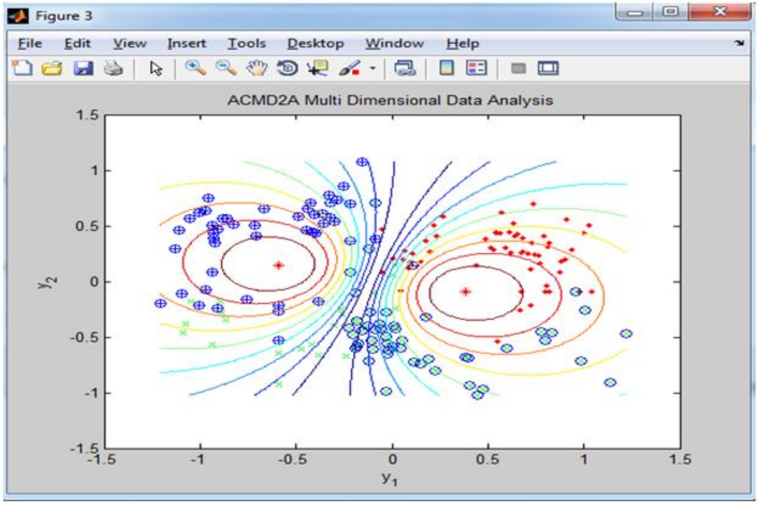
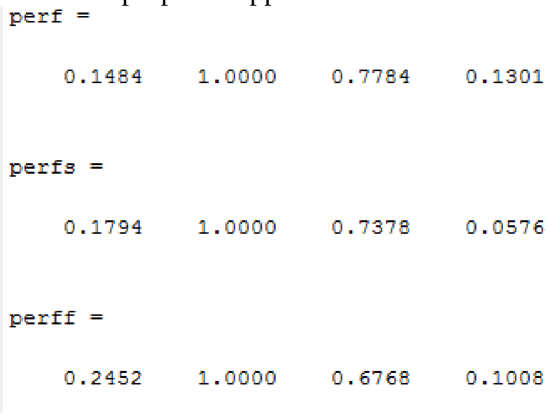


Fig.6. Multi Dimensional Data Analysis

The researcher ultimate aim to get the result in multi dimensional cluster base map reduce that is obtained from Mat lap simulation tool analyzer the dimensional vary every data set of machine learning process. The Command window of the proposed approach in MatLab is shown in figure 7.



```
perf =

    0.1484    1.0000    0.7784    0.1301


perfs =

    0.1794    1.0000    0.7378    0.0576


perff =

    0.2452    1.0000    0.6768    0.1008
```

Fig.7. Command Window Result analysis

DOI https://doi.org/2024-JICS-22490 | Generated on 2025-04-20 19:32:59

OPEN ACCESS

The result outcomes from command window perfs which denote the sub massive result in different map reduce the resultant iamge is shown in figure 8. The cluster base group separation is represented is shown in figure 9. The Cluster Base Result in different Data set is shown in figure 10.

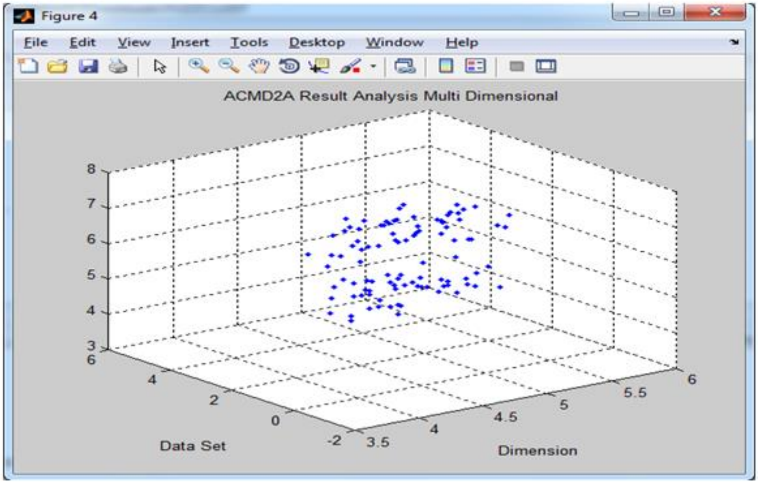


Figure 4
File Edit View Insert Tools Desktop Window Help
ACMD2A Result Analysis Multi Dimensional
8
7
6
5
4
3
6
4
2
0
-2
Data Set
3.5
4
4.5
5
5.5
6
Dimension

Fig.8. Result Analyses in Multi- Dimensional Data Set

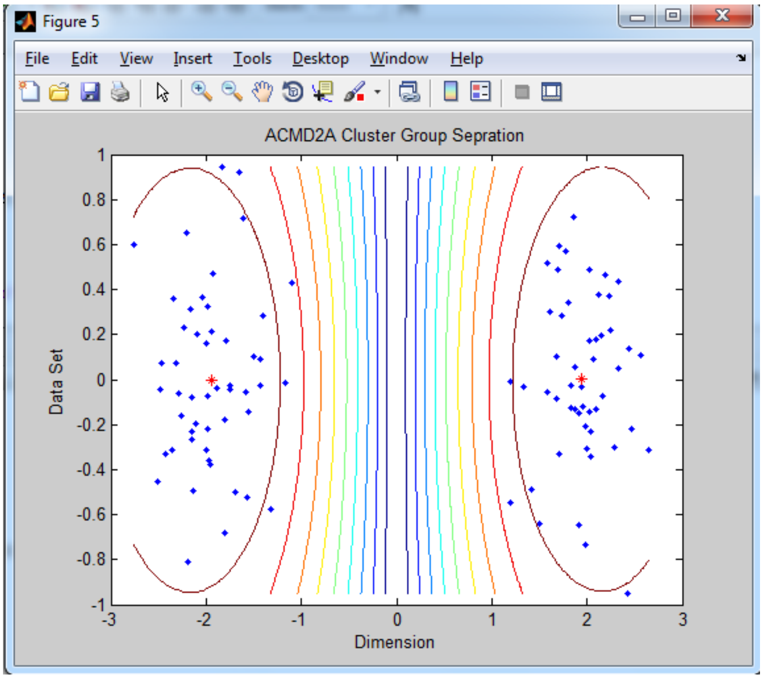
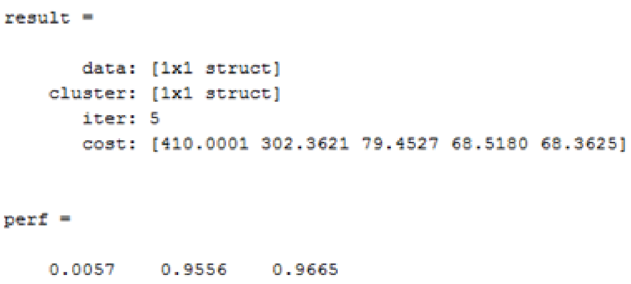


Figure 5
File Edit View Insert Tools Desktop Window Help
ACMD2A Cluster Group Sepration
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
Data Set
-3
-2
-1
0
1
2
3
Dimension

Fig.9. Cluster base Group Separation



```
result =

       data: [1x1 struct]
    cluster: [1x1 struct]
       iter: 5
       cost: [410.0001 302.3621 79.4527 68.5180 68.3625]


perf =

    0.0057    0.9556    0.9665
```

Fig.10. Cluster Base Result in different Data set

The figures 8 & 9 shows the final result of our proposed algorithm ACMDDA- Adaptive Cluster Multi Dimensional Data Analysis in Map Reduce Framework using Mat lab. The result is classifying cluster in reference to take multi dimensional data set in map reduce scenario. The graphical analysis of the proposed approach when it is implemented in Matlab is shown in figure 11.

DOI https://doi.org/2024-JICS-22490 | Generated on 2025-04-20 19:32:59
OPEN ACCESS

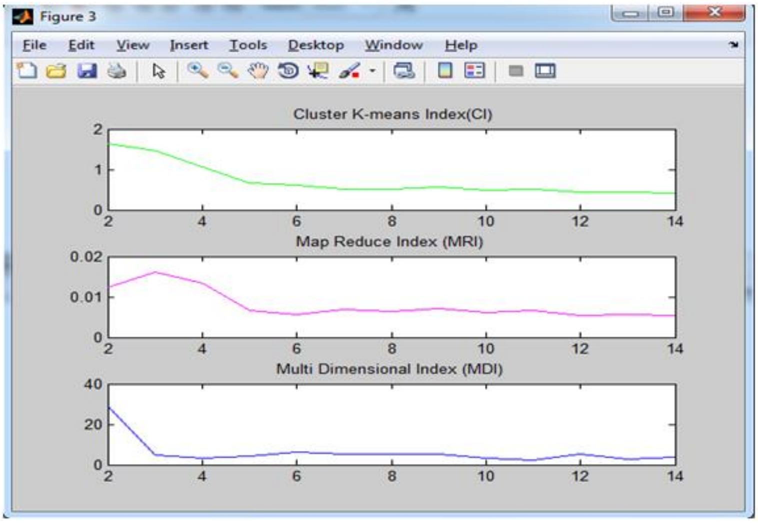


Figure 3
File Edit View Insert Tools Desktop Window Help
Cluster K-means Index(CI)
2
1
0
2 4 6 8 10 12 14
Map Reduce Index (MRI)
0.02
0.01
0
2 4 6 8 10 12 14
Multi Dimensional Index (MDI)
40
20
0
2 4 6 8 10 12 14

Fig.11. Graphical analyses in Mat lab.

4. Conclusion

The suggestion of this research is used to knob the huge quantity of Multi-Dimensional statistics by map diminish technique. Adaptive K-means Clustering algorithms are used to find the pattern in information and these algorithms must also scale well with the rising quantity of data. In this examine using k-means clustering algorithm can use to cluster the enormous dataset calculate the MSSE (Mean Sum Square Error) and find the accurateness, creation use dissimilar type of dataset in the Hadoop framework cluster presentation. Map Reduce manage the multi dimensional data divider and take on alike allowance on the portioned data. Adaptive K-means Cluster analysis based upon distance metric chosen and the decisive factor for formative the order of clustering. Depending upon the centroids may yield dissimilar results. The numbers of iterations necessary for cluster data can be concentrated by choose the right set of first set of centroids.

5. References

[1] DweepnaGarg, KhushbooTrivedi, B. B. Panchal, A Comparative study of Clustering Algorithms using MapReduce in Hadoop, International Journal of Engineering Research & Technology. 2.

[2] Prajesh P Anchalia, Anjan K Koundinya, Srinath N K, MapReduce Design of K-Means Clustering Algorithm, IEEE International Conference. (2013).

[3] Kashef, Shima, and Hossein Nezamabadipour, An advanced ACO algorithm for feature subset selection, Neurocomputing. 147: 271-279(2015).

[4] Fahim, A. M., A. M. Salem, F. A. Torkey, M. A. Ramadan , An efficient enhanced k-means clustering algorithm, Journal of Zhejiang University Science. 7(10):1626-1633(2006).

[5] Min Wei, Tommy WS Chow, and Rosa HM Chan, Mutual Information-Based Unsupervised Feature Transformation for Heterogeneous Feature Subset Selection, arXiv preprint arXiv. 1411: 6400 (2014)

[6] X. Zhang, C. Liu, S. Nepal, W. Dou and J. Chen, Privacy-preserving Layer over MapReduce on Cloud and Green Computing (CGC 2012), Xiangtan, China. 304-310(2012).

[7] X. Zhang, L. T. Yang, C. Liu and J. Chen, A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization using MapReduce on Cloud, IEEE Transactions on Parallel and Distributed Systems (TPDS). 25(2): 263-373(2014).

[8] Liu H. and Orban D., Cloud MapReduce: a MapReduce implementation on top of a cloud operating system, In: IEEE/ACM international symposium on cluster, cloud and grid computing. 464–474(2011).

[9] Arnab Roy, J. David Schaffer, Craig B.Laramee, New crossover operators for multiple subset selection tasks, arXiv preprint arXiv.1408: 1297 (2014).

[10] Slagter, Kenn, Ching-Hsien Hsu, Yeh-Ching Chung, and Daqiang Zhang, An improved partitioning mechanism

DOI https://doi.org/2024-JICS-22490 | Generated on 2025-04-20 19:32:59
OPEN ACCESS

for optimizing massive data analysis using MapReduce, The Journal of Supercomputing. 66(1): 539-555(2013).

[11] X. Zhang, C. Liu, S. Nepal, S. Pandey and J. Chen, A PrivacyLeakage Upper-Bound Constraint Based Approach for Cost-Effective Privacy Preserving of Intermediate Data Sets in Cloud, IEEE Trans. Parallel and Distributed Systems. (2012).

[12] E. Achtert, C. Böhm, H.-P. Kriegel, P. Kröger, A. Zimek, Robust, complete, and efficient correlation clustering, In SDM, USA. (2007).

[13] P. K. Agarwal, N. H. Mustafa, k-means projective clustering. 155–165(2004).

[14] C. Aggarwal, P. Yu, Redefining clustering for high-dimensional applications, IEEE TKDE. 14(2): 210–225(2002).

[15] Bosworth, Gray, Chaudhuri, Reichart, Pellow, Venkatrao, Data cube: a relational operator generalizing group by, cross-tab and subtotals,"Proc. 12th Int’l Conf. Data Eng, (ICDE). (1996).

[16] Deshpande, S. and R. Agarwal, Gupta, Naughton J., Sarawagi, Ramakrishnan, On the computation of multidimensional aggregates," Proc.22nd Int’l Conf. Very Large Data Bases (VLDB), (1996).

[17] Dean J., Ghemawat D. S, MapReduce: simplified data processing on large clusters, Commun ACM. 51: 107-113(2008).

[18] P. Jurczyk, L. Xiong, Distributed Anonymization: Achieving Privacy for Both Data Subjects and Data Providers, Proc. 23rdAnn. IFIP WG 11.3 Working Conf.Data and Applications Security XXIII (DBSec ’09). 191-207(2009).

[19] M. Deshpande, F. Naughton, Zhao, An array based algorithm for simultaneous multidimensional aggregates, In SIGMOD'97.

[20] Srivastava D., Ross, Fast computation of sparse data cubes, Proc.23rd Int'l Conf, Very Large Data Bases (VLDB). (1997).

[21] Han , Xin D., X. Li, W. B, Wah Starcubing: Computing iceberg cubes by top-down and bottom-up integration, In VLDB'03.

[22] Abouzeid, A., Bajda-Pawlikowski, K., Abadi, D.J., Rasin, A., Silberschatz, A, HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads. PVLDB 2(1)(2009).

[23] Chen, Z., Ordonez, C, Efficient OLAP with UDFs. Proc. Of DOLAP. (2008).

[24] G. Ananthanarayanan et al, Scarlett: coping with skewed content popularity in mapreduce clusters. (2011).

DOI https://doi.org/2024-JICS-22490 | Generated on 2025-04-20 19:32:59

OPEN ACCESS