An Intelligent Big Data Analytics System using Enhanced Map Reduce Techniques

S.Dhanasekaran, B.S.Murugan, V.Vasudevan

ABSTRACT – An Intelligent Big Data Analytics System using Enhanced Map Reduce Techniques include a set of Methods, applications and strategy which helps the organization and industry to bring together the data and information from outside sources and internal systems, as well as it is used to collect , classify, analysis and run the queries against the data and prepare the report for effective decision making. The Enhanced Map Reduced Techniques based on K-Nearest Neighbor (KNN) clustering Strategy works efficient as well as in an effective manner. We found that the existing MR – mafia sub space clustering Strategy have not performed effectively .Many clustering techniques are adopted in real world data analysis for example customer behavior analysis, medical data analysis, digital forensics, etc. The existing MR- mafia sub space clustering Strategy is inefficient because of continuously increase in the data size, and overlaying of the data blocks .The proposed KNN clustering Strategy mainly focused on the enhanced the Map Reduce techniques, and then to avoid the unnecessary input and output data, optimize the data storage in order to achieve the best out sourcing of data privacy. The proposed KNN clustering Strategy works effectively and that can be outsourced to cloud server.

Keywords: Big Data, Map Reduce, KNN clustering Strategy, Cloud Server, Subspace Clustering Strategy.

I. INTRODUCTION

Big Data Analytics techniques developed each and every day to fulfill the large number of customer's needs and necessity. It manipulate the large set of data set. The traditional computing techniques that cannot be processed efficiently for example: face book, you tube ,which contains the large data sets on every day which comes under the concept of Big Data .volume, variety ,velocity and also includes scale these are the Big Data features .The Big Data is one of the best method in the growing technology because it's to rectify the more byzantine problems .Map Reduce is one of the techniques comes under the Big Data methodology and its efficiently work on Big Data environment to dividing and partitioning the data and then to optimizing the storage capacity which ensued by clustering techniques .The clustering techniques rapidly used in many real world domains such as health care, social network ,image analysis and pattern recognition .

Revised Manuscript Received on December 15, 2019.

S.Dhanasekaran, Department of Computer Science and Engineering Kalasalingam Academy of Research and Education (Deemed To Be University)

B.S.Murugan, Department of Computer Science and Engineering Kalasalingam Academy of Research and Education (Deemed To Be University)

V.Vasudevan, Department of Computer Science and Engineering Kalasalingam Academy of Research and Education (Deemed To Be University)

The large scale data sets are to be efficiently managed by the clustering system. Grouping the similar kind of data is one of the major goals of the clustering techniques which can be easily processed to retrieve the data, begin protected and demonstrates the data and finally good trade off to the cloud storage.

Intention

The ultimate intention of this research work is to provide enhanced KNN clustering Strategy and to improve the capacity on cloud data set utilizing improved KNN clustering Strategy dependent on map reduce method for big data analytics.

Scope

The extent of our Research work is to improve the data accuracy and system security and also to improve the system reliability.

II. RELATED WORKS

Zhipenggao et al[1] proposed MR- mafia parallel sub space clustering Strategy. This Strategy handles large amount of multi- dimensional information then the author fully focused on MR- mafia sub space clustering Strategy it's based on map reduce information .the map reduce information's are partitioning and then tasks are parallelized because to perform a decent tradeoff between the disk access. The author obtained high capability and best application prospects of the proposed techniques .the more sub spaces are to be created so more noisy data's are to be accomplished in the clustering process the high dimensional data's are does not perform efficiently .it's very difficult to comparing other clustering Strategy the MR-mafia sub space clustering Strategy only performing a data partition does not providing security to all the data's.

Prajesh P Anchalia et al[2]focused on Map Reduce methods by making use of combiner then the Map Reduce to enhance the performance execution the combiner to perform the read and write operation between the number of median the single point failure cannot finding a solution of HDFS cluster (name node) these problems are to be occur.

BtissamZerharil et al [3] presented on the common view of big data Strategys, data mining clustering techniques, partition base clustering method, density base clustering Strategy, hierarchical clustering Strategy and Strategy challenges have also been discussed.

T.Mohanapriya et al[4] discovered Hadoop and Map Reduce structure is powerful benefits of this paper. Mainly to supporting a inter cluster resemblance and intra cluster resemblance further its provides best result in big data

environment. The major problems are computational complexity and scalability.

N.Vishnupriya et al[5] discovered on data mining techniques .the maximum advantages of the paper is to enhance ability and to be collect the huge information .the issues is focused on only to handle the 2 Dimensional and three Dimensional data sets.

Amar deep Kaur et al [6] proposed on sub space clustering techniques .This Strategy works efficiently to find sub space in high Dimensional data. the process of sub space clustering Strategy is computationally very expensive this is the major drawback.

Veronica et al [7] proposed on parallel K- means clustering Strategy the existing clustering techniques does not computing the cluster metrics .here the metrics are used for finding knowledge using big data datasets. The major issues of this Research work the big data's are to be clustered take a several time and we have not now deal with the metrics for measuring separation of the cluster.

Xunanliul et al[8] discovered on meta learning programmed with Map Reduce techniques the advantages are to reduce computational complexity and then to significantly producing smaller error rates. The problem is do not have the more capacity because to deal with large amount of data.

Weizhong Zhao et al[9] discovering parallel clustering Strategy based on map reduce method .this Strategy efficiently processing the large data sets using on commodity hardware .extremely high computational complexity and then providing a poor scalability this issues are accomplished.

Tanvir Habibsardars et al[10] using partition based clustering techniques to using the different datasets its perform efficiently and then to reducing the computational time .This Strategy only partition by the data and cannot providing security these are the major problems .

III.PROBLEM IN EXISTING SYSTEM

In existing framework MR-mafia parallel sub space clustering Strategy has been utilized which perform high dimensional spaces with the goal to lead the traditional clustering Strategy. The noisy data's are to be over lapping the high dimensional space then the execution assurances provides not only faithful provision of services to the cloud system. Yet also assets on in the home IT infrastructure .The sensitive information's are containing the datasets are used for clustering purpose eg: patient health information, commercial data, and behavioral data, etc directly outsourcing them, to public cloud server definitely to raises security concern.

IV. PROPOSED SYSTEM

Map reduce is one of the big data methodology in which we can optimize more extra space on large scale data set .the concept of map reduce is used to dividing a file into blocks and check for the block repeated in the storage .here the problem is arises s to verify the block is present or not in the storage.to overcome this problem to introduced by KNN clustering Strategy using map reduce techniques. the first step to perform uploading the trained data set each and every cluster which is relevant to medical information's after to performing KNN clustering Strategy to split file into number of chunks and hash code is generated to every chunks for the security purpose and finally the information's are store the cloud system, which activities can able to saves more time and increased the performance.

System architecture of proposed system

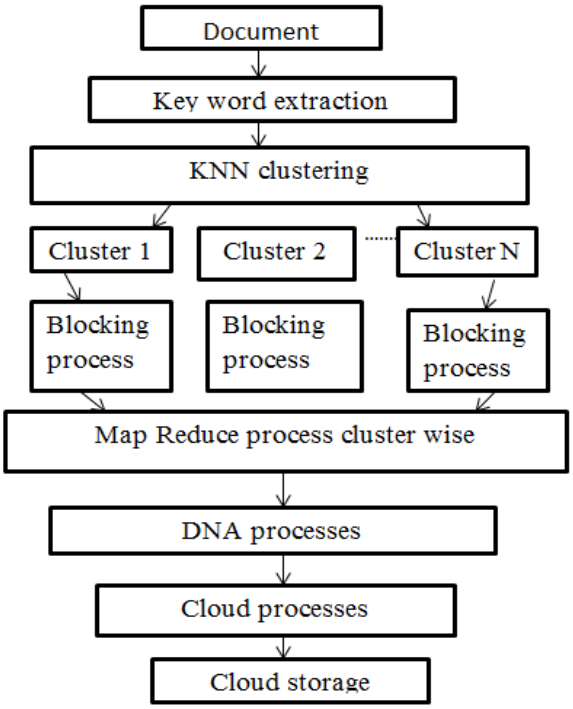


Fig 1: System architecture of proposed system

V. METHODOLOGY

Intelligent Big Data Analytics System methodology includes the following methodology. Map reduce is an innovative technology by which we can reduce more storage space on large scale dataset. The concept of map reduce is to divide a file into blocks and check for the block existence in the storage. If it is present no need to store the block.
Here the problem arises to verify the block is present or not on a huge number of blocks it will take more time. So the best way is to identify the file classification and search the block existence in particular cluster. Which saves more time and performance is increased.

- KNN Clustering Strategy
- Map reduce Strategy
- DNA (deoxyribonucleic acid) Strategy

KNN Clustering Strategy

Step 1: Get the File (F)

Step 2: Extract the keywords with weight age and store it in array K []

Step 3: Let N be the Number of Classification

Step 4: Initialize a Array Class _ Weight [N]

Step 5: Let M be the number of extracted Keywords

Step 6: For I = 1 to M

Step 7: Let K Word = k [I]

Step 7: For J = 1 to N

Step 8: Check the presence of K Word in J th Classification Keywords

Step 9: If it present Class _ Weight [J] = Class _ Weight [J] + K Word Weight

Step 10: Next J

Step 11: Next I

Step 12: Fetch the (Next) highest Class _ Weight Value and Index

Step 13: Add Index in Classification Array

Step 13: W = W + Fetched Class _ Weight

Step 14: if W >= Threshold then

Step 15: Go to Step 12

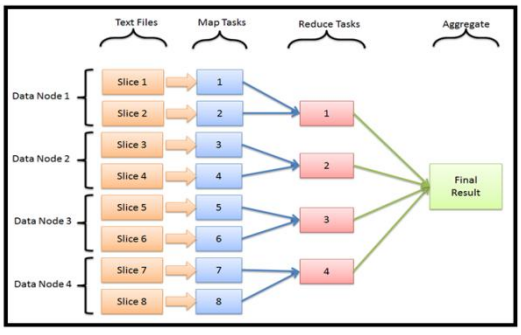
Step 16: Print all the categories in Classification Array

Step 17: Stop

Fig 2: KNN clustering Strategy

Map reduce Strategy

Map Reduce is one of the programming model is mainly used to computing process the Map Reduce Strategy easily organize the computation process. the huge amount of machines are to be run is the major advantages of the Map Reduce method.



Map reduce Strategy

Step 1: Start

Step 2: Read File (F) from the respective file.

Step3: Based on the Packet size file chunks (blocks) will be formed.

Step 4: For I=0 to N (total number of blocks)

Step 5: Generate the hash code for each chunks (blocks) using MD5 algorithm.

Step 6: Compare the hash code with existing hash code in the data base.

If exists get the id (block id) of the identical hash code for the LBA and

Increase the instance by Map Reducing Technique (Mapping to the

Existing block).

Else insert the new hash code to the database and get the id of the inserted hash code for the LBA Process and Do Step 7.

Step 7: Upload the block to the Cloud storage.

Step 8: End For

Step 9: Append all the blocks id of the file and create the Logical Block

Addressing (LBA) and maintain in the database.

Step 10: Stop

Fig 3: Map reduce Strategy

DNA Strategy

DNA is one of the programming languages of our genetic code. The information's are to be encodes and the major benefit of this Strategy to store a huge amount of data in a very small space. The data are built with block by block and then they included ID tags in each blocks.

Encryption process

Step 1: Get the Message

Step 2: Covert the String into the Streams

Step 3: Let Consider n be the length of String S1 (e.g. n = 7)

Step 4: Pad the beginning of each with a blank to simplify things (e.g. S1 = "_WRITERS")

Step 5: Fill an initially empty by 0

Step 6: Let M be a Original data Convert binary data to DNA sequences.

Step 7: Let M' = DNA Sequences

A=00,
T=01,
C=10, and
G=11.

Step 8: Apply the Base Pairing rule on M'

(A= 00, T= 01, C= 10, G= 11): M´= TAAT

Step 9: Applying complimentary rule M''

((AC) (CG) (GT) (TA)): M´´= ACCA

Step 10: Indexes: M´´´= 0706 **(Encrypted data)**

Fig 4: DNA Strategy for Encryption process

Decryption Process

Step 11: Convert numeric data to DNA sequences.

Step 12: M´´´=0706 (Input)

- By referring the DNA sequence: Sub-phase1 (Indexes): M´´= ACCA.
- By using Complementary rule: Subphase2((AC)(CG)(GT)(TA)):M´ = TAAT
- By using Base Pair Rule:

Sub-phase3 (A= 00, T= 01, C= 10, G= 11): M=01000001 (A) **(Output)**

Fig 5: DNA Strategy for Decryption process

VI.RESULT AND DISCUSSION

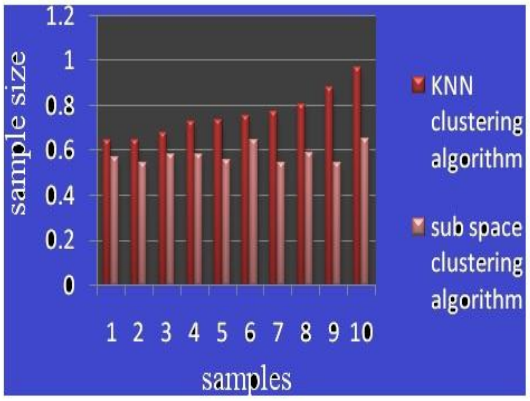


Fig:6 KNN clustering Strategy and subspace clustering Strategy performance analysis

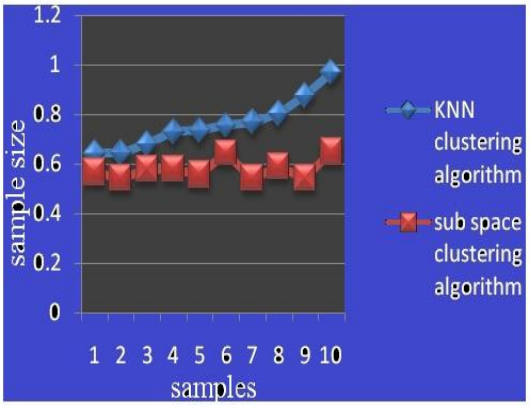


Fig 7: KNN clustering Strategy and subspace clustering Strategy Accuracy analysis.

Table1 I: Dataset

Data points	1	2	3	4	5
Average map time	12	30	48	55	70
Shortest map time	1	3	4	4	6
Shuffle task time	50	99	130	167	190
Average reducer time	2	2	1	2	1
Total time for map reduce	65	120	170	213	270
Number of spilled records	11	11	11	11	11
Number of killed task	0	0	0	1	1
Time taken to coverage	350	590	840	1080	1300

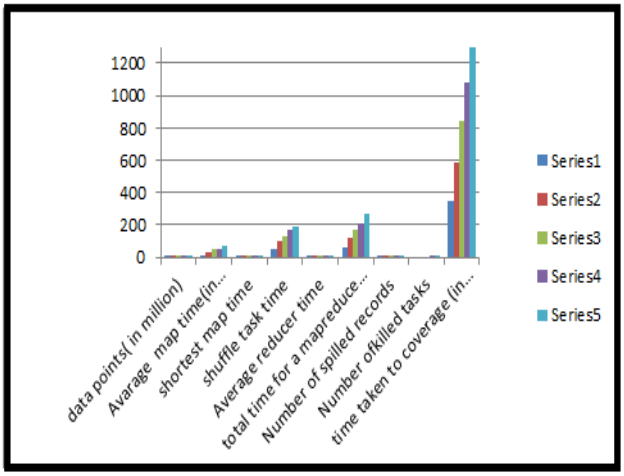


Fig:8 Map Reduce performance analysis

Accuracy comparison table

Table II: KNN clustering Strategy and MR-mafia subspace clustering Strategy accuracy comparison.

Strategy	simulated datasets (d = 500, Ns = 500)	Real datasets (d = 50, Ns = 7680)
K-means cluster	0.90	0.85
Subspace cluster	0.86	0.71

Performance Evaluation

The graph and tables are shows performance and accuracy analysis of KNN clustering Strategy and sub space clustering Strategy. The KNN clustering Strategy to be perform efficiently and to manage the clustering techniques .KNN clustering Strategy to avoiding the noisy data and then to optimizing the communication cost between nodes .moreover to achieve the independent calculation and also to performing load balance in each node and hash code is generated in each blocks is to providing the secure data transmission because the de-duplication concept is to gives the high speed of transmission then the KNN clustering Strategy using map reduce concept to overcome the sub space clustering Strategy difficulties. Then the KNN clustering Strategy to working efficiently.

VII. CONCLUSION AND FUTURE ENHANCEMENT

This Research work efficiently presented on KNN clustering Strategy based map reduce techniques for big data analytics. This Research work fully focused on KNN clustering and map reduce techniques are involved to overcome the sub space clustering problems. This intelligent system is to associated with centroid to the each data points and assign each cluster center randomly and then choose the most corresponding data points. Finally every computation processes are organized by the map reduce techniques. In future , more powerful Hash code generation techniques may be included to provide authentication and advanced compressing concept will be used for quick data uploading process in the cloud storage.

REFERENCES

1. MR-Mafia: Parallel Subspace Clustering Strategy Based on Map Reduce for Large Multidimensional Data sets, ZhipengGao, YidanFan, and 2018 IEEE.
2. Dhanasekaran.S and Vasudevan.V. Multiple Intelligent Agent Coordination Strategy for Categorizing and Searching Appropriate Cloud Services, IEEE Xplore Digital Library, (2018), 387-391.
3. S.Dhanasekaran, Dr.V.Vasudevan, "A Dynamic Multi-Intelligent Agent System for Enhancing the Cloud Service Negotiation", International Journal of Applied Engineering Research, vol. 10, no. 43, pp. 30469-30473, 2015.
4. Dhanasekaran.S & Vasudevan.V., A Smart Logical Multi agent System for Consolidating Suitable Cloud Services, International Journal of Computer Science and Information Security,**14** (9) (2016), 517-522.
5. Data Clustering using Map Reduce for Multidimensional Datasets, N.Vishnupriya1, Dr.F.Sagayaraj Francis2, International Advanced Research Journal in Science, Engineering and Technology,Vol.2,Issue8,DOI10.17148/IARJSET. 2015. 2810, August 2015,
6. A novel Strategy for fast and scalable subspace clustering o f high-dimensional data, Amar deepKau r * and AmitavaDatta, Journal of BigData, 2017, springer.
7. Dhanasekaran.S et al., De-noising of images from salt and pepper noise using Hybrid Filter, Fuzzy Logic Noise Detector and Genetic Optimization Algorithm (HFGOA), Multimedia Tools and Applications Springer,**78** (318) (2019), pp 1-17.
8. Dhanasekaran.S et al., "Brain Tumor Segmentation Using Convolutional Neural Networks In MRI Images", Journal of Medical Systems, Springer, 43 (9) (2019), pp 1-17.
9. Parallel K-Means Clustering Based on MapReduce, Weizhong Zhao1,2, Huifang Ma1,2, and Qing He1, M.G. Jaatun, G. Zhao, and C. Rong (Eds.): CloudCom 2009, LNCS 5931, pp. 674–679, 2009. c Springer-Verlag Berlin Heidelberg 2009.
10. Partition based clustering of large datasets using map reduce framework an analysis of recent themes and direction Tanvir Habib sardar,ZahidAnsari,http:// www.jourals.elsevier.com/future computing and informatics journal 3 2018.
11. Enhanced Map Reduce Techniques for Big Data Analytics Based on K- Means Clustering, S.Dhanasekaran, B.S.Murugan and V.Vasudevan, IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing – INCOS19, 2019.