Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 9, September -2017

Improved Efficiency of NoSql using Optimized K-Mode

Sandeep Kaur[1], Er. Gurleen Kaur Dhaliwal[2]

[1]*Research Scholar, Department of Computer Science and Engineering, Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib*
[2]*Assistant Professor, Department of Computer Science and Engineering, Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib*

Abstract-*Data Mining is the process to extract unknown, valid patterns and relationships that provide useful information. Non relational databases are a broad class of database management systems identified by non-adherence to the widely used relational database management system model. Non relational databases are not built fundamentally on tables, and generally do not use SQL for data manipulation. Non relational database systems are often highly optimized for retrieval and appending operations and often offer little functionality beyond record storage. The reduced run-time flexibility compared to full SQL systems is compensated by marked gains in scalability and performance for certain data models Clustering technique defines classes and put objects which are related to them in one class on the other hand in classification objects are placed in predefined classes. There are various algorithms used for clustering. These are k means algorithm, k-medoid algorithm, k-nearest neighbour algorithm, k-mode algorithm etc. K-Modes algorithm is an extension of K-Means and uses simple matching dissimilarity function instead of Euclidean distance. The major drawback of k-mode is that the user needs to define the centroid points. The nature-inspired harmonic algorithm is hybridized to optimize the k-mode algorithm. Harmonic K-Mode Algorithm is proposed in this work that reduces the computation time and improves the execution time for cluster generation.*

Keywords- *Data Mining, Clustering, K-Mode, Harmonic Search.*

I. INTRODUCTION

Non relational databases are broad class of database management systems identified by non-adherence to the widely used relational database management system model. Non relational databases are not built primarily on tables, and generally do not use SQL for data manipulation. Non relational database systems are often highly optimized for retrieval and appending operations and often offer little functionality beyond record storage. The reduced run-time flexibility compared to full SQL systems is compensated by marked gains in scalability and performance for certain data models.

In Non Relational databases, the CAP theorem states that it is impossible for a distributed computer system to simultaneously provide all of consistency, availability, and partition tolerance.

Consistency :-If and how a system is in a consistent state after the execution of an operation. A distributed system is typically considered to be consistent if after an update operation of some author all readers see his updates in some mutual data source.

- Availability :- That a system is designed and implemented in a way that allows it to continue operation (i. e. allowing read and write operations) if e. g. nodes in a cluster crash or some hardware or programming parts are down due to upgrades.
- Partition Tolerance :- The ability of the system to continue operation in the presence of network partitions. These happen if atleast two “islands” of network nodes arise which (temporarily or permanently) cannot connect to each other. Some people also understand partition tolerance as the ability of a system to cope with the dynamic addition and removal of nodes (e. g. In maintenance purposes; removed and again included nodes are considered an own network partition).

Given such limitations, there was a strong push towards an approach that considers the trade-offs among different aspects of the distributed system and design them in a way that is most effective for different applications. In Non relational database, BASE transactions are used instead of ACID transactions (atomicity, consistency, isolation, durability - four obvious features of traditional relational database systems in sql).
BASE transactions are used in nosql databases which means:-

- Basically Available(B)-an application works basically all the time.
- Soft state(S)-Does not have to be consistent all the time.
- Eventually Consistent(E)-But will be in some known-state eventually.

II. CLUSTERING ALGORITHM

Clustering is the technique which defines classes and put objects in one class which are related to them which means to put the objects into one group having similar properties and objects having dissimilar properties into another group. It has alienated the large data set into groups or clusters according to similarity in properties [7]. It is a form of unsupervised learning that means how data should be group the data objects (similar types) together will be not known in advance. Clustering technique defines classes and put objects which are related to them in one class on the other hand in classification objects are placed in predefined classes [9]. The clustering algorithms can be generally classified into five types:

- Hierarchical Clustering Method
- Partitioning Based Clustering Method
- Density Based Clustering Method
- Grid Based Clustering Method
- Model Based Clustering Method

Hierarchical Clustering Method: In this method, Hierarchical breakdown of the given set of data objects is created. It can be defined into two approaches agglomerative and divisive. Agglomerative approach is the bottom up approach. This approach starts with each object forming a separate group. It combines groups close to one another until all the groups are merged into one group. Divisive approach is top down approach and cluster is dividing into smaller clusters until each object is in one cluster [5, 7].

Partitioning Based Clustering Method: The general criteria of partioning, is combining the high similarity of the samples inside clusters with high dissimilarity between separate clusters. Most of the partitioning process is distance based process. Given k, the number of partitions to create, it takes an initial partitioning and then uses an technique of iterative relocation that attempts to improve the partitioning by transferring objects from one group to another. The partitioning clustering algorithms are k-mean, Medoids Method, PAM, CLARA etc.

Density Based Clustering Method: Generally in partitioning methods, cluster objects are based on distance between objects. While Spherical shaped clusters can be discovered by density based method and come across difficulty in inventing Clusters of random and arbitrary shapes. In Arbitrary shapes new methods are used called as density-based methods which are based on the notion of density. DBSCAN is an example of density based clustering method [5].

Grid Based Clustering Method: The grid-based clustering approach uses a multi resolution grid data structure. It quantizes the object space into finite number of cells that form a grid structure on which all of the operations for clustering are performed. The main advantage of this approach is its fast processing time, which is typically independent of the number of data objects, yet dependent on only the number of cells in each measurement in the quantized space.

Model Based Clustering Method: In Model based method each of the clusters is best fitted to the given model. It may locate clusters by constructing a density function that reflects the space distribution of the data points [9].

III. CLUSTERING USING K-MODE

Clustering is the technique which defines classes and put objects which are related to them in one class on the other hand in classification objects are placed in predefined classes. Clustering means to put the objects which have similar properties into one group and objects having dissimilar properties into another group [7]. Clustering has alienated the large data set into groups or clusters according to similarity in properties. Outliers are the data points which are present outside the clusters. It uses sophisticated data analysis tools and visualization techniques to segment the data and evaluate the probability of future events[8].
K-Mode algorithm is an extension of K-Means which is the partitioning based clustering algorithm. It is an eminent algorithm for clustering data set with categorical attributes and is famous for its simplicity and speed. Modes are used to represent centroids instead of Mean values and finally a frequency based method is used to find centroids in each iteration of the algorithm [6].

Algorithm for K-Mode Clustering:

1. *Generate K clusters by arbitrarily selecting data objects and choose K initial cluster centre, one for every of the cluster.*
2. *Assign data object to the cluster whose cluster centre is near toward it according to equation*

$d(X,Y) = \sum_{K=1}^{m} \delta(x_i, y_i)$ ..... *Eq (1)*

$where \delta(x_i, y_i) = \begin{cases} 0, & x_i = y_i \\ 1, & otherwise \end{cases}$ .....*Eq (2)*

3. Update the K cluster base on allocation of data objects. Calculate K latest modes of every one clusters.
4. Repeat step 2 to 3 awaiting no data object has changed cluster relationship otherwise some additional predefined criterion is fulfil.

IV. LITERATURE SURVEY

Ravi Sankar Sangam and Hari Om [1] discussed the limitations of distance function used in this algorithm with an illustrative example and then proposed a similarity coefficient based on Information Entropy. Clustering is the process of organizing dataset into isolated groups such that data points in the same are more similar and data points of different groups are more dissimilar. The k-modes algorithm well known for its simplicity is a popular partitioning algorithm for clustering categorical data. This paper analyzed the time complexity of the k-modes algorithm with proposed similarity coefficient. The main advantage of this coefficient is that it improves the clustering accuracy while retaining scalability of the k-modes algorithm and performs the scalability tests on synthetic datasets.

Parneet Kaur and Kamaljit Kaur[2] gave comparative study of clustering techniques and addresses benefits and limitations of clustering techniques. Data mining is a set of problem solving skills, instructions and methods applied upon variety of domains to discover and create useful systems that are used to solve practical problems. Clustering technique defines classes and put objects which are related to them in one class on the other hand in classification objects are placed in predefined classes.

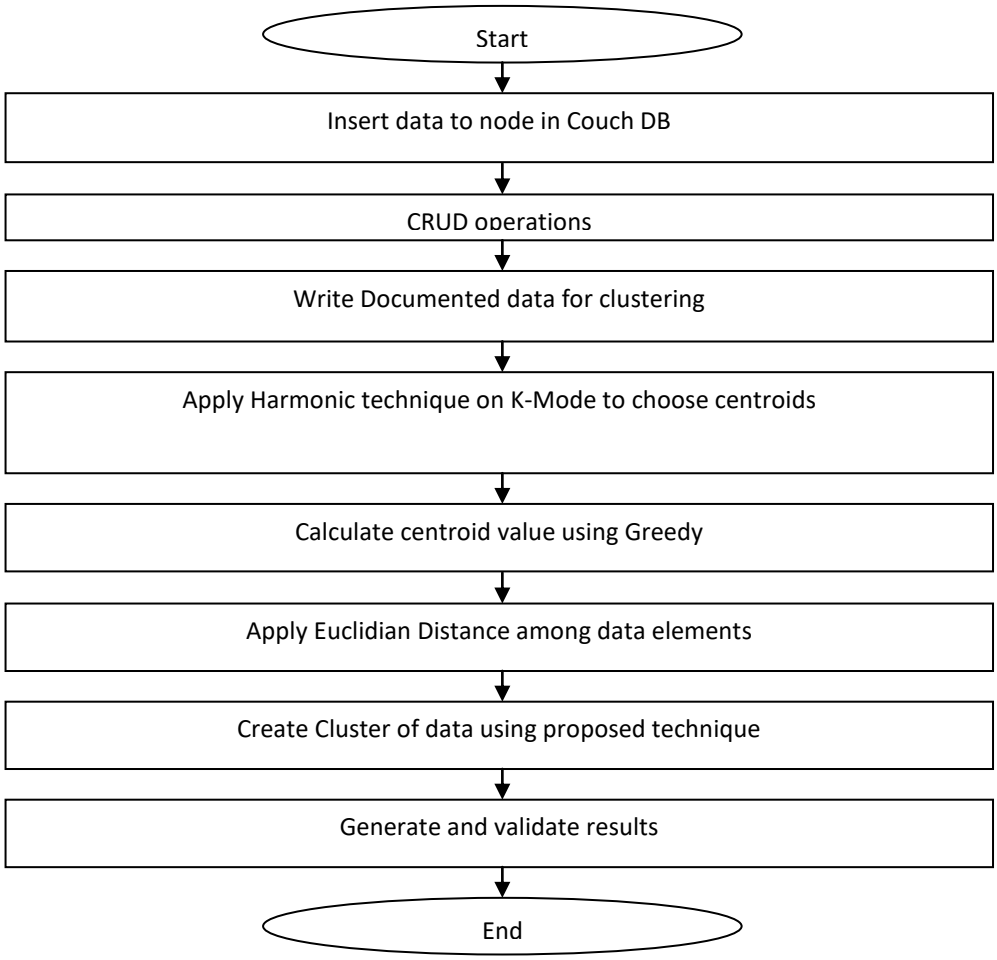
Abraham Jaison [3] defined that World of computing is witnessing an immense growth in all aspects. Increased usage of internet for communication and the growth of Social Networking led to the invention of NoSQL database technologies capable of handling large amount of structured or unstructured data. Apache Cassandra is such an open source distributed data management system designed to handle large amount of data and is now being widely deployed in production environments. One of the important factors developers look for is cost effective deployment of Cassandra clusters by improving the resource utilization of existing clusters without increasing the number of nodes. In the production environment, only one instance of Cassandra can be deployed on one node of a cluster which can lead to underutilized resources, if high specification servers are used. The cost involved increases as we increase the number of nodes in the cluster. In this paper, we propose a method to deploy an M+n node Cassandra cluster on M servers using Docker containers which allows deployment of multiple Cassandra instances on a single system and prove how effectively it can improve the performance of Cassandra NoSQL database with improved resource utilization without scaling the database into more number of nodes.

Preeti Arora et al. [4] defined that the two most popular clustering algorithms K-Means and K-Medoids are evaluated on dataset transaction10k of KEEL. The input to these algorithms are randomly distributed data points and based on their similarity clusters has been generated. The comparison results show that time taken in cluster head selection and space complexity of overlapping of cluster is much better in K-Medoids than K-Means. Also K-Medoids is better in terms of execution time, non sensitive to outliers and reduces noise as compared to K-Means as it minimizes the sum of dissimilarities of data objects.

V. PROBLEM FORMULATION

Non relational databases are not built fundamentally on tables, and generally do not use SQL for data manipulation. Non relational database systems are often highly optimized for retrieval and appending operations and often offer little functionality beyond record storage. The reduced run-time flexibility compared to full SQL systems is compensated by marked gains in scalability and performance for certain data models. Non relational database management systems are useful when working with a huge quantity of data when the data's nature does not require a relational model. The data can be structured, but Non relational database is used when what really matters is the ability to store and retrieve great quantities of data, not the relationships between the elements.
Clustering categorical data is a complex task since there is no natural order among the categorical values. The k-modes algorithm is a popular clustering algorithm in this regard since it is linearly scalable with respect to the dataset size.
K-Modes algorithm which creates appropriate number of clusters without need of prior input of number of clusters, K. The results obtained are found to be better than the original K-Modes in terms of the quality of clusters. The current algorithm has generated clusters such that the similarity of objects in a cluster in terms of number of attributes with matching values is maximum. This way this algorithm can obtain a cluster of objects form the given dataset with maximum similarity without providing the value of K initially.
After this technique an attempt may be made to reduce the computation time for clusters generation, especially with a large number of attributes. Also an algorithm for mixed datasets with no dependency on K as an input will be proposed.

Flow Chart



VI. RESULTS AND DISCUSSION

Execution Time: Execution time means the time taken required by the computer to perform a given set of computations. If Execution time is less than the clustering algorithm is better than algorithms having more Execution time.

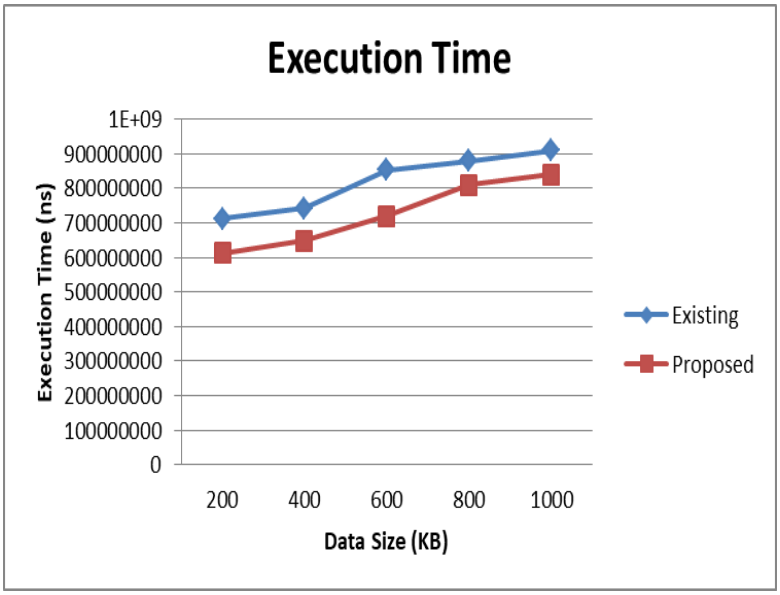


Figure 1: Execution Time

Figure1 show the performance degradation of Item retrieval when the database size increases in the number of Items. Retrieving and execution of an Item in Couch DB on a 1000KB Item database takes on average over 900000000 ns. There is a little degradation in performance of proposed scheme as we increase the data size. Whereas previous algorithm shows the performance degradation of item retrieval when the database size increases in the number of Items in terms of

time. Retrieving an Item in Couch DB on a at 200KB data size it takes 700000000 ns unit of time. There is a little degradation in performance of proposed scheme as we increase the time duration.

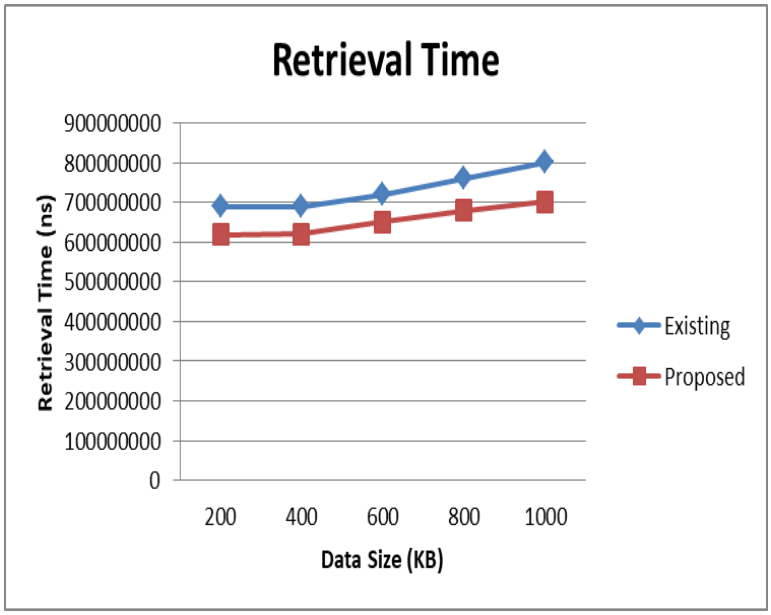


Figure 2: Retrieval Time

Figure 2 show the Item retrieval when the number of datases are increases in the network. Retrieving an Item in Couch DB on a 200KB Item database takes on average over 600000000 ns. There is a little degradation in performance of proposed scheme as we increase the number of database. Whereas previous algorithm shows the performance degradation of Item retrieval when the data size increase.

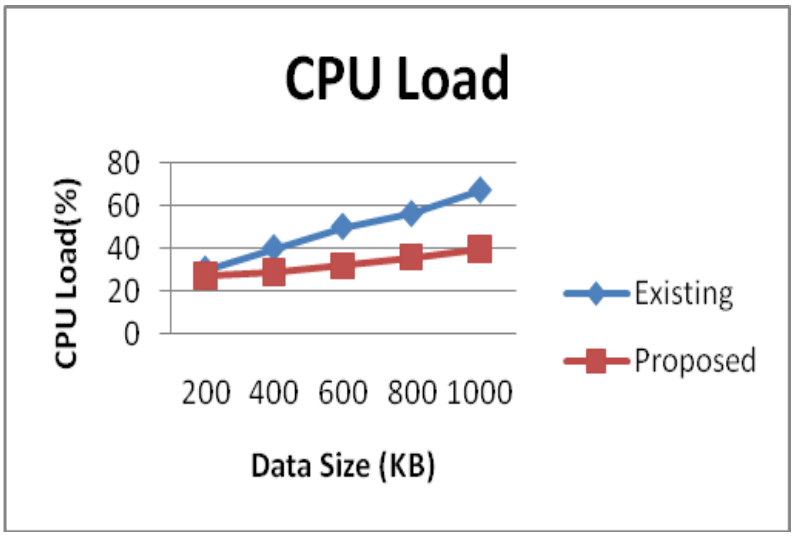


Figure 3: CPU Load

Figure 3 shows the performance degradation of in terms of load as the number of database increased gradually. When the number of databases time taken to search a particular item in the database also increased. But as we compare our proposed scheme with the previous algorithm we can clearly say that the performance of our scheme is far better than that of previous algorithm.

VII. CONCLUSION

As in a context management platform the number of context publications always keeps rising, further data partitioning schemes will also be researched, to better handle the always-increasing data load. The NoSql ecosystem, unlike relational databases, is headed towards specialization, so different solutions are headed in different directions, leaving the door open for new players to emerge, and making the ecosystem an exciting and ever-evolving field. For the future scope we can implement the distributed database on all nosql databases. When the number of database time taken to search a particular item in the database also increased. Retrieving an Item in Couch DB on a 1000KB Item database takes on average over 35 ms as compare to that of 50 KB in 30 ms. There is a little degradation in performance of proposed scheme as we increase the data size.

REFERENCES

1. Ravi Sankar Sangam, Hari Om, "The k-modes algorithm with entropy based similarity coefficient", 2nd International Symposium on Big Data and Cloud Computing, Procedia Computer Science, Volume: 50, 2015, pp: 93-98
2. Parneet Kaur, Kamaljit Kaur, "Clustering Techniques in Data Mining For Improving Software Architecture: A Review", International Journal of Computer Applications, ISSN: 0975 – 8887, Volume: 139, No.9, April 2016, pp: 35-39
3. Abraham Jaison, Kavitha N , "Docker for optimzation of cassandra NoSql deployments on node limited clusters", International Conference on Emerging Technological Trends, ISSN: 0975–8887, Volume: 135, No.7, February 2016, pp: 22-24
4. Preeti Arora, Deepali, Shipra Varshney, "Analysis of K-Means and K-Medoids Algorithm For Big Data", International Conference on Information Security & Privacy, Volume: 78, 2016, pp: 507-512
5. Bharati M., Ramageri, 2010 "Data Mining Techniques and Applications": Indian Journal of Computer Science and Engineering, Vol. 1, pp. 301-305.
6. Bhardwaj, B., 2016 "Text Mining, its Utilities, Challenges and Clustering Techniques", International Journal of Computer Applications, ISSN: 0975–8887, Volume: 135, No.7, pp: 22-24
7. Bijuraj, L.V., 2013 "Clustering and its Applications", Proceedings of National Conference on New Horizons in IT – NCNHIT, India, pp. 169- 172.
8. Gupta, V., 2013 "A survey on Data Mining: Tools, Techniques, Applications, Trends and Issues", International Journal of Scientific & Engineering Research, Vol. 4, Issue. 3.
9. Marir, F., Said, H., Al-Obeidat, F., 2016 "Mining the Web and Literature to Discover New Knowledge about Diabetes", The 3rd International Workshop on Machine Learning and Data Mining for Sensor Networks, Elsevier, Vol. 83, pp. 1256-1261.

