ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.7

An Implementation of Hierarchical Clustering Algorithm in (DCC) Environment

C. Beaulah
M. phil Scholar
Department of Computer Science
Annai Vailankanni Arts and Science College, Thanjavur, Tamil Nadu, India

Abstract:
Nowadays increase in worldwide business lead to offices dispersed across geographical location .Hence data be loosely distributed across regionalized large scale databases across regionalized offices. To perform data mining it is required to merge distributed data and perform data mining algorithm on it. Cloud computing poses a diversity of challenges in data mining operation arising out of the dynamic structure of data distribution as against the use of typical database scenarios in conventional architecture. This article presents a method to implement Hierarchical Agglomerative Clustering Algorithm into such way so as to make it right for large dataset and increase its efficiency by executing task in parallel. The result shows that with increase in data set linear growth of execution time.

Index Terms: Star cluster, hierarchal agglomerative clustering, virtual k mean, cloud computing.

I. INTRODUCTION

With the rapid development of information technologies, internet begins to pervade our daily life and has become into a new life style that enriches people's living contents. SE (search engine), a crucial part of the internet, is an important tool for us to acquire information. In searching process, how to find and download pages that are most relevant to users' topics has now become the key for the topical search engine. At present, there are two common types of searching strategies .The first is a content evaluation-based searching strategy such as Fish-Search and Best-First The content describes topics accurately and thus the relevancy between them can be calculated well and truly. This type of strategy, however, ignores structural information of links. Hence, it has disadvantages when forecasting the accuracy of link values and database. An article we will focus on Hierarchical Agglomerative bottom up merging fashion based algorithm and suit it to Geographical distributed data set. Our aim is to increase the efficiency of agglomerative clustering algorithm as well as to make it suit for large data. To implement this we require the cloud computing virtualized environment Virtualization is a key technology used in data centers to optimize resource. Assume data distributed among different node. By virtualization we create instances of each geographical distributed node. The article consists of five sections: Section I provides introduction on cloud computing, Hierarchical Agglomerative Clustering and Virtualization concept. Section II describe the design of modified agglomerative clustering technique along with algorithm that suit used for cloud platform. Section III describes the experimental setup to implement on cloud based architecture. Section IV provides us with experimental results and benefits on implementing it. Section V describes the conclusion and future work to be performed.

II. LITERATURE SURVEY:

S. Pippal, [8] Cloud computing is a computing paradigm where services and data reside in common space in scalable data centers, which are accessible via authentication. Cloud computing [1] services can form a strong infrastructural and service foundation framework to provide any kind of service oriented computing environment. Ad-hoc clouds [2, 3] enable existing infrastructure as cloud compliant, the available resources in the environment are utilized non-intrusively. An Ad-hoc cloud is very efficient solution to problems faced by organizations to Venture into remote areas. Education-cloud, where a cloud computing framework is harnessed to manage Information system of an Educational institution would be highly efficient in terms of accessibility, manageability, scalability and availability. An ad-hoc cloud would enable us harness services offered by Fixed Education-cloud and services created and composed within ad-hoc cloud. The ad hoc cloud as derives data and cloud service from fixed cloud, further they are connected using an ad hoc link (V-SAT). The S, P and V nodes in the ad hoc data center represents Super-node (Permanent node at remote location with ad hoc connectivity with the fixed cloud to facilitate cloud configuration at remote site), Persistent-node (organization's hosting cloud with data services) and Volunteer-nodes (other participate nodes in an group).

T. R. G. Nair, **[7]** Increase in the usage of cloud computing has sparked a new interest among researchers of data mining. Using contemporary algorithms has proven to be inefficient on the cloud. It is not suited for large and highly distributed database because the time for execution is very large. [1] Cloud have emerged as a computing infrastructure to enables rapid delivery of computing resources as a utility in a dynamically scalable virtualized manner [2]. Data mining is a process of discovering meaningful patterns and relationships that are hidden in large data set [3]. Simply stated data mining refers to extracting or "mining" knowledge from large amounts of data [4]. Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects. The cluster is collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Though there are many algorithms

which takes care of large database but huge memory usage is always a concern. Using cloud to process and store database can solve this problem as it can take care of more memory requirement very easily.

III. DESIGN OF EFFICIENT AGGLOMERATIVE CLUSTERING TECHNIQUES

It have been argued to perform effectively on large databases, the algorithm must require no more than one scan of the database, have the ability to provide "best " answer so far, be suspend able, stoppable and resumable, be able to update the results incrementally etc. Keeping these points in mind the basic idea would be to read the subsets of database, apply clustering algorithm and combine the results with those from prior samples and proceed in this way till all the data is available in main cluster. A hierarchical clustering algorithm is suitable for small dataset but for making it suite to large dataset. We will divide it in two tasks - 1. Micro clustering stage 2. Macro clustering stage. As shown in Fig. 1. Modified Hierarchical Agglomerative clustering performs processing at three layers.

A. Apply Virtual K Mean

Layer 1: In this layer data from various geographical distributed dataset are loaded into individual virtualized node. We apply virtual k-mean algorithm on every node which resolution form k number of cluster on individual node. This output will be stored on separate file created at individual node. Thus macro clustering occurs at this layer.

B. Merging Files

Layer 2: In this layer the outputted files which consist of K-centriod and cluster are merging into single file called Master file. To reduce any error normalization is performed on this master file. Thus master file contain data which are cluster analysis and outlier error free. D and E represent individual cluster while modified algorithm represents central centroid of data cluster generated by k mean algorithm. Thus this modification provides us with following benefits:

1) We will be able to use Hierarchical Agglomerative clustering algorithm for large set of data.

2) The efficiency of the algorithm has been increase due to performing macro clustering on large data set followed by micro clustering on outputted centroid of data cluster.

3) Parallelisms of task reduce the time required for execution.

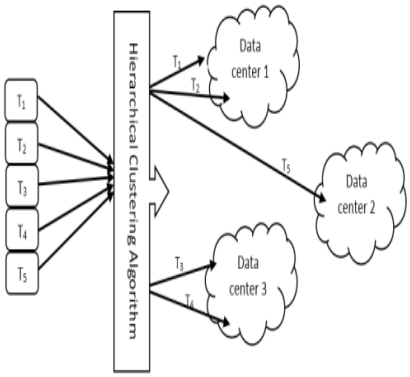

Figure.1. Cloud data center

Algorithm

Step 1: Define the number of nodes N which would be equal to geographical distributed data-set.

Step 2: Apply k-mean clustering algorithm on each node individually.

Step 3: The output will be each file consists of k number of centroid and respective cluster stored in separate files.

Step 4: Perform merging of separated files in to single master file. Thus single master file consist of (k * n) of centroid.

Step 5: Perform normalization on master file to reduce error by outlier and cluster analysis.

Step 6: Apply Hierarchical Agglomerative Clustering algorithm on master file which will output dendogram as shown in below Fig. 2.

Thus in Fig. 2 A, B, C, D and E represent centroid and data cluster represent the cluster formed by grouping data objects using k mean algorithm.

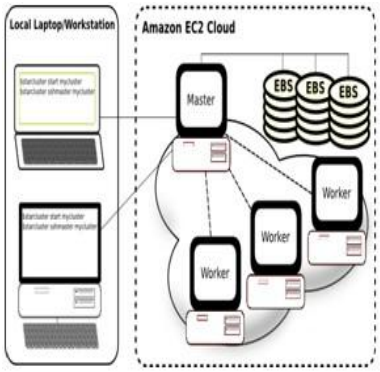

Figure.2. Logical structure

IV. EXPERIMENTAL SETUP

The implementation of the over concept in cloud architecture require master and slave node. Used for master and slave architecture we contain used MIT's Star Cluster platform whose logical structure is shown in Fig. 3. Star Cluster creates Amazon EC2 instance for master and all the nodes. It enables SSH access between them the memory blocks between them are shared by NFS. The nodes have MySQL and Java pre-installed. The data sets have been stored in MySQL and the modified algorithm has been written in JAVA. On the master node, we execute commands on the nodes by using SSH to ensure that the commands are run in a parallel manner. We pass the commands to the qs Perform merging of separated files in to single master file. Thus single master file consist of (k * n) of centroid.

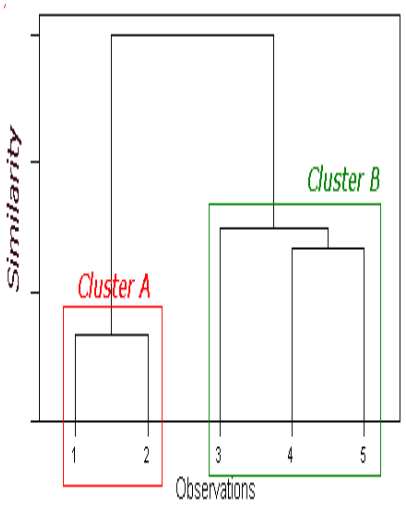

Figure.3. Den-dogram

The main difference between the normal

Algorithm and modified is as shown in output figure 2. The

HAC has A, B Grid Engine. The k-mean algorithm gets executed on each slave node. Thus the task required to be executed by layer 1 shown in Fig. 1 gets executed at slave node. The result of virtual k-mean from each node is transferred to the Master Node. After normalizing the result HAC algorithm is run on the master to obtain the final results. The sample algorithm for executing Hierarchical Agglomerative clustering is shown in Fig. 4. Thus task required to be executed by layer 2 and layer 3 shown in Fig. 1 gets executed at master node.

```
Input: A dataset D
Output: A hierarchy tree of clusters
Allocate each centroid (k-mean result) as o(i) in D as a single cluster;
Let C be the set of the clusters;
While |C|>1 do
        For all clusters X, Y ? C do
                Compute the between-cluster similiarity S(X, Y),
        end
        Z=X ? Y, where S(X, Y) is minimum;
        Remove X and Y from C;
        C=C ? Z;
end
```

Figure. 4. Sample code for HAC algorithm

V. RESULT ANALYSIS

We contain applied the adapted HAC algorithm on sample medicine database have attributes weight with ph .To compare its efficiency we contain executed by variation of nodes. The total number of data mined be number of nodes*number of rows in every node. Because nodes increase the data also increases but doesn't deteriorate the performance as shown in. present be a linear increase in instant required for execution. With quadratic increase in information across cloud environment the time necessary for execution increases linearly. Therefore efficiency of the modified algorithm have been increase greatly by parallelism of task on cloud base architecture.

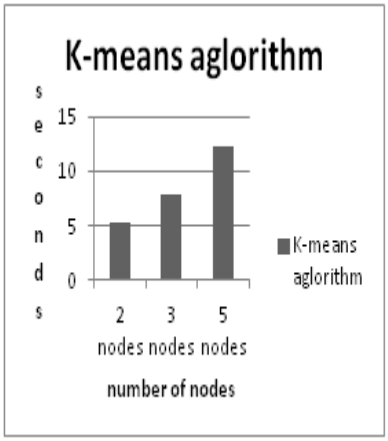

Figure. 5. Result analysis.

VI. CONCLUSION AND FUTURE WORK

Cloud is a highly dispersed computing model which is having inherent advantages of scalability, availability, elasticity, pay per use etc. Thus modifying algorithm to suit cloud architecture can enables above benefits. In addition to this it provides us with following benefits like Hierarchical Agglomerative clustering can handle large dataset, increase efficiency of algorithm and parallelism has reduce time required for execution. Thus the result shows that cloud architecture is providing additional ad-vantages for data mining. In future we can compare the results obtained from cloud plat-form with map reduce framework to understand the effectiveness.

VII. REFERENCES

[1]. Z. X. Hou, X. S. Zhou, J. H. Gu, Y. L. Wang, and T. H. Zhao, "ASAAS: Application software as a service for high per-formance cloud computing," in Proc. of 2010 12th IEEE International Conference on High Performance Computing and Communications (HPCC), pp. 156-163, 2010.

[2]. W. T. Tsai, X. Sun, and J. B. Sooriya, "Service oriented cloud computing architecture," in Proc. IEEE 2010 Seventh International Conference on Information Technology.

[3]. U. Fayyad, G. P. Shapiro, and P. Symth, "From data mining to knowledge discovery in databases," 0738-4602-19196, A Magzine37-53.

[4]. Data Mining and Analytics Resources. [on line] Available: http://www.kdnuggets.com/gpspubs/aimagkdd-overview-1996-Fayya d.pdf

[5]. J. W. Han and M. Kamber. Data Mining Con-cepts and Techniques. Morgan Kaufmann Publishers, Champaign CS497JH, fall 2001. [Online]. Available: http:// www.cs. uiuc.edu/~hanj/bk2/

[6]. Logistic Regression and Newton's Method. [Online]. Available: http://www.stat.cmu.edu/~cshalizi/350/ lec-tures/08 /lecture-08.pdf

[7]. T. R. G. Nair and K. L. Madurai, "Data min-ing using K-mean integrating data fragment in cloud environment," IEEE.\

[8]. S. Pippal, V. Sharma, S. Mishra, and D. S. Kushwaha, "Secure and efficient multitenant database for an ad hoc cloud," Securing Ser-vices on the Cloud (IWSSC), pp. 46-50, 2011.

[9]. J. Z. Wang, J. G. Wan, Z. Liu, and P. Wang, "Data mining of mass storage based on cloud computing," in Proc. of 2010 9th International Conference, Grid and Cooperative Computing (GCC), pp. 426-431, 2010.

[10]. M. Comerio, H.-L. Truong, and C. Batini, "Dustdar, S, cloud service engineering; Ser-vice-oriented computing and applications (SOCA)," 2010 IEEE International Confe-rence on Digital Object Identifier, pp.1-6, 2010.