ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

An Efficient Hierarchical Agglomerative Clustering Algorithm in Distributed Cloud Computing Environment

R.Anitha [1], Dr.M.Mayilvaganan [2]

Research Scholar, Department of Computer Science, PSG College of Arts and Science, Civil Aerodrome Post, Coimbatore, India[1]

Associate Professor, Department of Computer Science, PSG College of Arts and Science, Civil Aerodrome Post, Coimbatore, India[2]

ABSTRACT: This paper describes about how clustering algorithms are used in cloud computing. Cloud computing is a big shift from the traditional way business think about IT resources. Today's cloud computing technology has been emerged to manage large data sets efficiently and due to rapid growth of data. To perform data mining it is required to merge distributed data and perform data mining algorithm on it. This document describes about How hierarchical Agglomerative algorithm used for large dataset and increase efficiency by executing task in parallel the result shows that with increase in data set linear growth of execution time and finding the efficiency of the execution time between the hierarchical clustering methods.

KEYWORDS: Cloud Computing, Hierarchical Agglomerative Clustering, virtual k mean, star cluster.

I. INTRODUCTION

Cloud Computing is a very growing technology now a days so that data mining is the most important thing and must be used in cloud computing for security, efficiency and productivity and less time consuming. Using contemporary algorithms on cloud has proven to be lack on efficiency. It does not support for large and distributed database because the time for execution is very large. Cloud has emerged as a computing infrastructure that enables rapid delivery of computing resources as a utility in a dynamically scalable virtualized manner. Data mining has emerged as a way for identifying pattern and trends from large quantities of data. Simply stated data mining refers to extracting or "mining" knowledge from large amounts of data. Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other group.

A hierarchical clustering method works by grouping data objects into tree of clusters. Hierarchical clustering method can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up (merging) or top-down(merging) fashion.

In this paper we will focus on Hierarchical Agglomerative bottom up merging based algorithm and suit it for 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 centres to optimize resource. Assume data distributed among different node. By virtualization we create instances of each geographical distributed node. This bottom- up strategy starts by placing each object in

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

its own cluster and then merges these atomic clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied.

II.LITERATURE REVIEW

Today's rapid technological development has lead to exponential growth of human needs, many aspect of human life increase dependent with technological change, as in the field of medical, financial even up to education. This fact requires HLI has to perform continuous efforts to adapt information from industries, thus graduates meet the needs of industry. Both educators and employers have recognized the importance of educating the professionals that design, develop, and deploy information systems [5].

Therefore, we can use data mining to recognize the relationship between educational institutions and industries, data mining application in education also called the Educational Data Mining. Educational Data Mining (EDM) is the process of converting data from educational systems to useful information that can use by educational software developers, students, teachers, parents, and other educational researchers [6]. Many study discuss EDM, also job opportunities in the industry for college graduates, such as data mining for needy of students based on improved RFM model [6], educating experienced IT professionals by addressing industry's needs [5], jobs for young university graduates [8], measure the relationship between training and job performance [8].

Henchman and Archibald talk about meeting the needs of the mining engineering sector Error! Reference source not found., extension education to meet market demand [11], training, job satisfaction, and workplace performance

[11]. Also development and application of metrics to analyze job data [13], international competitiveness, job creation and job destruction Error! Reference source not found, undergraduate career choice [15], constructing the search for a job in academia [16], students and their performance [16].

This research focuses on industrial training of the final-year students, the objective of industrial training is to experience and understand real life situation in industry and their related environment. Another objective is adequate to practice theory learned in solving real problems, in various aspects such as concept planning, design, construction and administration projects, focus on organizational structure, industrial training and identify the role of certain positions in the industry. Industrial training will perform professionalism and ethics. Clustering can use to explore information about real career world, job requirement for organization structure, business operation and administration functions.

Clustering is concerned with grouping together objects that are similar to each other and dissimilar to the objects belonging to other clusters [18]. Clustered is used to group items that seem to fall naturally together [19]. Various type of clustering: hierarchical (nested) versus partitioned (un-nested), exclusive versus overlapping
versus fuzzy, and complete versus partial [20]. Clustering is an unsupervised learning process that partitions data such that similar data items grouped together in sets referred to as clusters, this activity is important for condensing and identifying patterns in data [21].

Clustering techniques apply when there is no class to predict but rather when the instances divide into natural groups. These clusters presumably reflect some mechanism at work in the domain from which instances are drawn a mechanism that causes some instances to bear a stronger resemblance to each other than they do to the remaining instances. Clustering naturally requires different techniques to the classification and association learning methods we have considered so far [19].

Hierarchical clustering creates a hierarchy of clusters, which may represent in a tree structure called a dendogram. The root of the tree consists of a single cluster containing all observations, and the leaves correspond to individual observations. Algorithms for hierarchical clustering are generally either agglomerative, in which one starts at the leaves

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

and successively merges clusters together; or divisive, in which one starts at the root and recursively splits the clusters. Any valid metric use as a measure of similarity between pairs of observations. The choice of which clusters to merge or split is determined by a linkage criterion, which is a function of the pair wise distances between observations [wikipedia].

Agglomerative clustering starts with N clusters, each of which includes exactly one data point. A series of merge operations then followed, that eventually forces all objects into the same group [22]. Hierarchical algorithms find successive clusters using previously established clusters. These algorithms can be either agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

III. AN EFFICIENT MODEL OF AGGLOMERATIVE CLUSTERING TECHNIQUES

It has been argued that to perform effectively on large databases, the algorithm does not require the database scan more than once. It can able to provide "best "answer so far, be suspendable, stoppable and resemble. The hierarchical clustering algorithm is mostly used for small dataset but in this paper we are making it suite to large dataset. We will divide in to two tasks

- Microclustering stage
- Macroclustering stage.

As shown in Fig 1. Hierarchical Agglomerative Clustering will perform three layers in cloud computing.

A. Apply Virtual K Mean

Layer 1: In this layer data from a various geographical distributed dataset are loaded into individual virtualized node. Then apply virtual k mean algorithm on each node it will form a K number of cluster on individual node. This output will be stored on separate file created at individual node. Thus macroclustering occurs at this layer

B. Merging Files

Layer2: In this layer the stored output files which consist of k-centroid and cluster are merging into single file called Master file. If any error or normalization problem means that will be solved on this master file alone. This master file will contain the data which are cluster analysis and outlier analysis free.

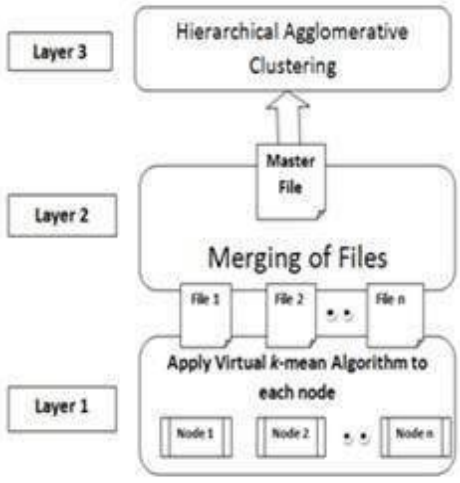

Fig. 1. Design

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

C. Hierarchical Agglomerative Clustering

Layer 3: In this layer on the outputted master file apply Hierarchical agglomerative clustering algorithm. The output will be in the form of dendogram macroclustering will be occurs in this layer.

D. Algorithm

Step 1: Define the number of nodes N it must be equal to geographical distributed dataset.

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

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

Step 4: Now Merge 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.

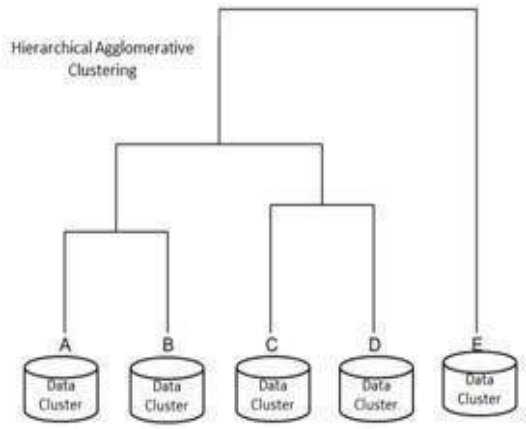

Fig 2 : dendogram

The main difference between the normal algorithm and modified is as shown in output figure2. The HAC has A,B,C,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 macroclustering on large data set followed by microclustering on outputed centroid of data cluster.

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

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

IV. IMPLEMENTATION OF HAC

The implementation of the above concept in cloud architecture requires master and slave node. For master and slave architecture we have used MIT's StarCluster platform whose logical structure is shown in Fig. 3. StarCluster 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 sub of Sun Grid Engine.

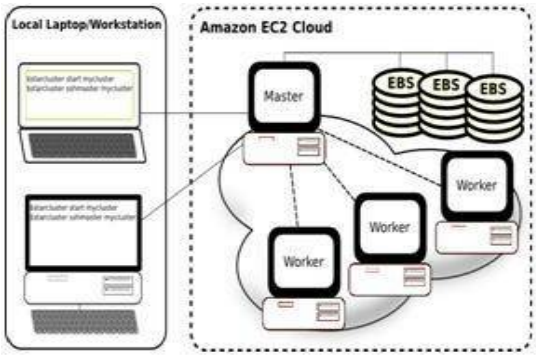

Fig. 3. Logical Structure

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.

```
SIMPLEHAC(d_1, ..., d_N)
 1  for n ← 1 to N
 2  do for i ← 1 to N
 3     do C[n][i] ← SIM(d_n, d_i)
 4     I[n] ← 1  (keeps track of active clusters)
 5  A ← []  (assembles clustering as a sequence of merges)
 6  for k ← 1 to N − 1
 7  do (i, m) ← arg max_{(i,m): i≠m ∧ I[i]=1 ∧ I[m]=1} C[i][m]
 8     A.APPEND((i, m))  (store merge)
 9     for j ← 1 to N
10     do C[i][j] ← SIM(i, m, j)
11        C[j][i] ← SIM(i, m, j)
12     I[m] ← 0  (deactivate cluster)
13  return A
```

Fig. 4. Sample code for HAC algorithm

V.RESULT ANALYSIS

We have applied the modified HAC algorithm on sample market database having attributes weight and ph. To compare its efficiency we have executed with variation of nodes. The total number of data mined are number of nodes*number of rows in each node. Thus as nodes increase the data also increases but doesn't deteriorate the

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

performance as shown in Fig. 5. There is a linear increase in time required for execution. With quadratic increase in data across cloud environment the time required for execution increases linearly. Hence efficiency of the modified algorithm has been increase greatly by parallelism of task on cloud based architecture.

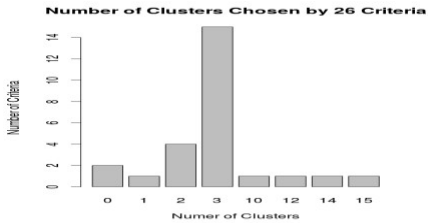

Fig. 5. Result analysis

VI.CONCLUSION AND FUTURE WORK

Cloud computing often referred to as simply "the cloud" is the delivery of on-demand computing resources. It has the advantages of scalability, reliability, availability, pay per use and etc. Modifying Agglomerative Clustering algorithm is used to handle a large dataset as well as increase the efficiency of the algorithm as well as take less time for an execution process So that the HAC algorithm is used for cloud computing to get a data efficiency. In future purpose we can compare the cloud computing with map reduce framework to get the best effectiveness.

REFERENCES

[1] Eldesoky, A.E, M.Saleh, N.A. Sakr, "Novel Similarity Measure for Document Clustering Based on Topic Phrase", International Conference on Networking and Media Convergence (ICNM 2009), 24-25 March 2009, p. 92-96, [DOI>10.1109/ICNM.2009.4907196]
[2] Myatt, Glenn J, "Making Sense of Data II", 2009, Wiley, Canada
[3] Maimon, Oded, Lior Rokach, "Decomposition Meth-odology For Knowledge Discovery And Data Min-ing", World Scientific, 2005, [DOI>10.1.1.150.2473]
[4] Lu, Qifeng, Yao Liang, "Multiresolution Learning on Neural Network Classifier : A Systematic Approach", International Conference on Network-Based Information Systems, 2009, p.505-511, [DOI> 10.1109/NBiS.2009.83]
[5] Bin, Deng, Shao Peiji, Zhao Dan, "Data Mining for Needy Students Identify Based on Improved RFM Model: A Case Study of University," International Conference on Information Management, Innovation Management and Industrial Engineering (ICIII), 2008, vol. 1, pp.244-247, [DOI> 10.1109/ICIII.2008.128]
[6] Cardoso, Ana Rute, "Jobs for young university graduates", Economics Letters, Volume 94, Issue 2, February 2007, p.271-277, [DOI>10.1016/j.econlet.2006.06.042]
[7] Devaraj, Sarv, S. Ramesh Babu, "How To Measure The Relationship Between Training and Job Perform-ance", Communications of the ACM, Volume 47, Is-sue 5, p.62-67, 2004, [DOI> 10.1145/986213.986214]
[8] Heuchan, Archibald, J.F., "Industry and Higher Education—Meeting the Needs of the Mining", Engineering Sector 105th Annual General Meeting ofthe Canadian Mining, Metallurgy and Petroleum, Montreal, Quebec, 2003
[9] Hsia, Tai Chang, An Jin Shie, Li Chen Chen, "Course planning of extension education to meet market de-mand by using data mining technique – an example of Chinkuo technology university in Taiwan", 2008, [DOI>10.1016/j.eswa.2006.09.025]
[10] Jones, Melanie K, Richard K. Jones, Paul L.Latreille, Peter J.Sloane, "Training, Job Satisfaction and Work-place Performance in Britain: Evidence from WERS 2004", LABOUR, Volume 23, p.139–175, 2009, [DOI>10.1111/j.1467-9914.2008.00434.x]
[11] Hsia, Tai Chang, An Jin Shie, Li Chen Chen, "Course planning of extension education to meet market de-mand by using data mining technique – an example of Chinkuo technology university in Taiwan", 2008, [DOI>10.1016/j.eswa.2006.09.025]
[12] Jones, Melanie K, Richard K. Jones, Paul L.Latreille, Peter J.Sloane, "Training, Job Satisfaction and Work-place Performance in Britain: Evidence from WERS 2004", LABOUR, Volume 23, p.139–175, 2009, [DOI>10.1111/j.1467-9914.2008.00434.x]