Scientific Journal of Impact Factor(SJIF): 3.134
e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406
International Journal of Advance Engineering and Research Development
Volume 1,Issue 12, December -2014

ENHANCING CLUSTERING OF CLOUD DATASETS USING IMPROVED AGGLOMERATIVE ALGORITHMS

Mrs. Parekh Madhuri Harsh[1], Prof.Jay M Jagani[2]

[1]*Computer Engineering, Student, Darshan Institute Engineering & Technology, Rajkot, Gujarat, India*
madhurisuchak18@gmail.com
[2] *Computer Engineering, Prof, Darshan Institute Engineering & Technology, Rajkot, Gujarat, India*
jay.jagani009@gmail.com

Abstract:- *Cloud computing is the latest technology that delivers computing resources as a service such as infrastructure, storage, application development platforms, software on Internet-based. Cloud computing is focus on delivery reliable, secure, fault tolerant, sustainable and scalable infrastructures for hosting on Internet Based application services. On basis of infrastructure service huge amount of data is stored in the cloud from distributed nodes which needs retrieved very efficiently. In Cloud Computing using of Clustering Process from Heterogeneous Network fetch the data. Hierarchical clustering is group data over a variety of scales by creating a cluster tree or dendrogram. The retrieval of information from cloud takes a lot of time as the data is not stored in an organized way. Data mining is thus important in cloud computing. So integrate data mining and cloud computing which will provide agility and quick access to the technology. The integration should be so strong that it will be able to deal with increasing production of data and will help in efficient mining of massive amount of data. In this dissertation work we provide brief description about cloud computing and clustering techniques. This dissertation work proposes a model that applies move traditional improved Agglomerative Hierarchical Clustering Algorithms on Heterogeneous Network.*

Keywords:- *Cloud Computing, Clustering, Hierarchical Algorithm, Agglomerative Algorithm, Distributed Algorithm, Hadoop.*

I. INTRODUCTION

1.1 Overview of Cloud Computing.

Cloud is designed to be available everywhere, all the time. By using redundancy and geo-replication, cloud is so designed that services be available even during hardware a failure including full data center failures. Cloud computing is anything involves delivering hosted services over the internet. It is a paradigm in which information is permanently stored in server on the internet and stored temporarily on client.[1] It is work with large groups of remote servers are networked which allow centralized data storage and online access to computer services or resources. Cloud resources are usually not only shared by multiple users but are also dynamically reallocated per demand.

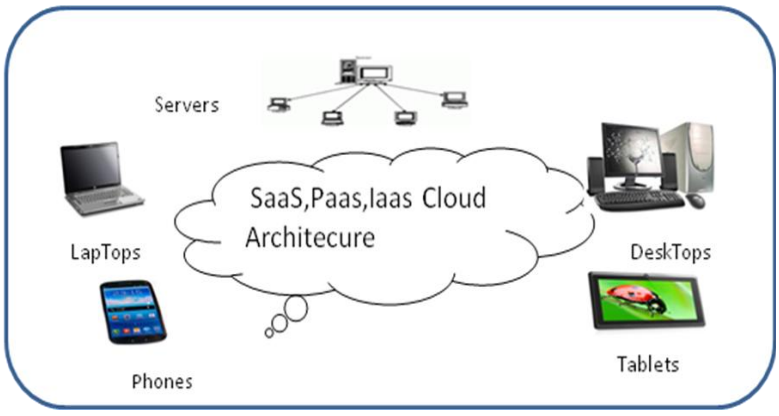


Figure-1. Cloud Computing Logical Diagram

1.2 Advantages and Disadvantages of Cloud Computing

1.2.1 Advantages of Cloud Computing[2]

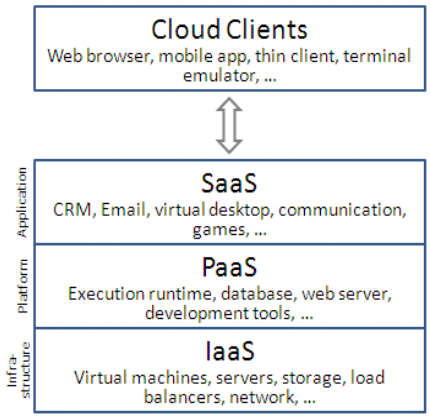
1. Device & Location is independence.
2. Increase flexibility.
3. Increase Security.
4. Increase Storage.
5. Decrease Cost.
6. Speed and Scalability.
7. Easy Access Information.
8. Automatic Software Integration.

1.2.2 Disadvantages of Cloud Computing[2]

1. Possible Downtime.
2. Lack of control.
3. Management Capabilities.
4. Reliability and Availability
5. Appropriate clustering and Fail over
    a. Data Replication
    b. System monitoring (Transactions monitoring, logs monitoring and others)
    c. Maintenance (Runtime Governance)
    d. Disaster recovery
    e. Capacity and performance management
6. Lock In.
7. Regulatory and Compliance Restrictions.

1.3 Cloud Computing Architecture

Cloud computing system can be divided into 2 sections, one is Front end and another one is Backend. The front end is the interface for the user example client and the back end is the cloud section for the whole system. Front end and Backend connected with each other via network like internet.
There are 3 types of layers related to Cloud services.



Figuare – 2 Cloud Architecure

Cloud Infrastructure Services (Infrastructure as a Service "IaaS"): This service provider bears all the cost of servers, networking equipment, storage, and back-ups.Rather than purchasing servers, software, data-center space or network equipment, clients instead buy those resources as a fully outsourced service.[3]

Cloud Application Services (Software as a Service "SaaS"): This service provider will give your users the service of using their software, especially any type of applications software. Google Apps., Salesforce.com,and various other online applications are use cloud computing as Software-As-Service (SAAS) model.[3]
Cloud Platform Services (Platform as a Service "PaaS"): Platform cloud services are used by software developers to build new applications and by operations managers to manage their application, compute and storage cloud services.[3]

II BACKGROUND WORK & RELATED WORK

2.1 Overview of Clustering

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. It is useful technique for the discovery of data distribution and patterns the underlying data.

2.2 Types of Clustering

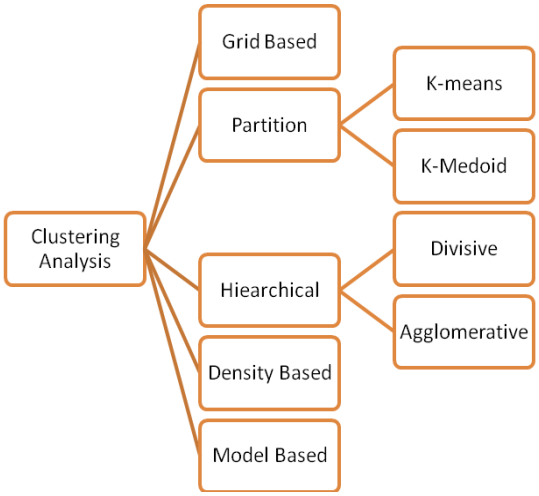


Figure- 3. Types of Clustering

Agglomerative and Division
There are two basic approaches to generating a hierarchical clustering:

Agglomerative: Start with the points as individual clusters and, at each step, merge the closest pair of clusters. This requires defining the notion of cluster proximity.

Divisive: Start with one, all-inclusive cluster and, at each step, split a cluster until only singleton clusters of individual points remain. In this case, we need to decide which cluster to split at each step.

Basic Agglomerative Hierarchical Clustering Algorithm
Many hierarchical agglomerative techniques can be expressed by the following algorithm, which is known as the Lance-Williams algorithm.

2.3 Implementation of Background Work

1) **Layer-1 Apply Virtual K Mean**
   a. Data fetch from various geographical distributed dataset are loaded into individual virtualized node.
   b. Then we apply virtual k-mean algorithm on each node which will form k number of cluster on individual node. This output will be stored on separate file created at individual node.
   c. Thus Macro clustering occurs at this layer. [4]
2) **Layer-2 Merging File**
   a. The outputted files which consist of k- centriod and cluster are merging into single file called Master file.
   b. To reduce any error normalization is performed on this master file. Thus master file contain data which are cluster analysis and outlier error free.[4]
3) **Layer-3 Hierarchical Agglomerative Clustering(HAC)**
   a. Apply Basic Hierarchical Agglomerative Clustering algorithm on outputted master file.
   b. The output in dendogram.
   c. Thus Micro-clustering occur at this layer.[4]

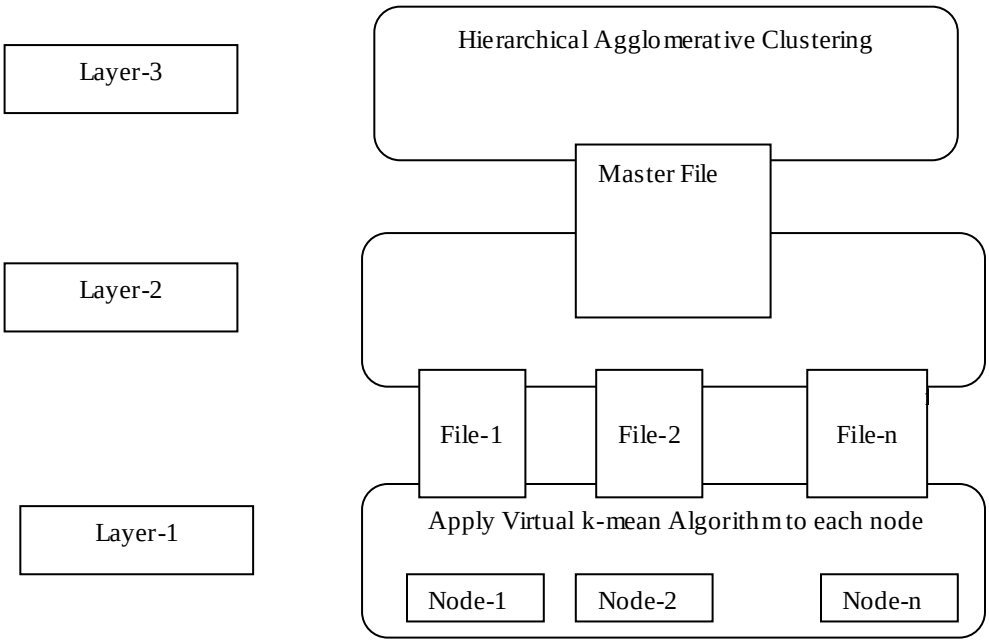


Figure - 4 Modified Hierarchical Algorithms

2.4 Limitation Of Background Work

Most k-means-type algorithms require the number of clusters - - to be specified in advance, which is considered to be one of the biggest drawbacks of these algorithms. 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, Required that efficiency of algorithm has been increase greatly by parallelism of tasks by Hadoop architecture.

2.5 Related Work

The Related Work in these field :- Esha Sarkar, C.H Sekhar[5] Proposed it provides fast access to data, provides the statistics of usage of cloud storage space, scalability and helps in mining large data sets which are heterogeneous in nature. Manpreet Kaur1, Mrs.Sukhpreet Kaur2[6] Proposed Clusters with different sizes in the tree can be valuable for discovery. It provides nearest distance between data points. Bhagyashree Ambulkar, Vaishali Borkar[7] Proposed Cloud computing allows the users to retrieve meaningful information from virtually integrated data warehouse that reduces the costs of infrastructure and storage .Hui Gao, *Jun Jiang, Li She, Yan F.*[8]Proposed Map Reduce framework can apply in large- scale dataset clustering satisfactory. B.Thirumala Rao, N.V.Sridevi,V.Krishna Reddy, L.S.S.Reddy[9] Proposed Hadoop lacks performance in heterogeneous clusters where the nodes have different computing capacity. In this paper we address the issues that affect the performance of hadoop in heterogeneous clusters. *Ms. E. Suganya, Mr. S. Thiruvenkatasamy* [10] ProposedThe Main goal managing effectively the technology needs the institution such as delivery of software, providing development platform, storage data and computing.

III PROPOSED WORK

3.1 Overview of Proposed Work

Different Agglomerative Hierarchical clustering algorithm has advantages over each other. Improved Performance of Agglomerative Hierarchical Clustering Algorithm is by Hadoop Tool for large set of data. The efficiency of the algorithm is increase. Parallelism of tasks reduces the time required for execution.Use CURE Agglomerative Hierarchical Clustering Algorithm.

3.2 CURE Agglomerative

Clustering Using Representatives

1) Draw a random sample from the data set.
2) Partition the sample into *p equal sized partitions.*
3) Cluster the points in each cluster using the hierarchical clustering algorithm to obtain m/pq clusters in each partition and a total of m/q clusters.
4) Eliminate outliers. This is the second phase of outlier elimination.
5) Assign all data to the nearest cluster to obtain a complete clustering .

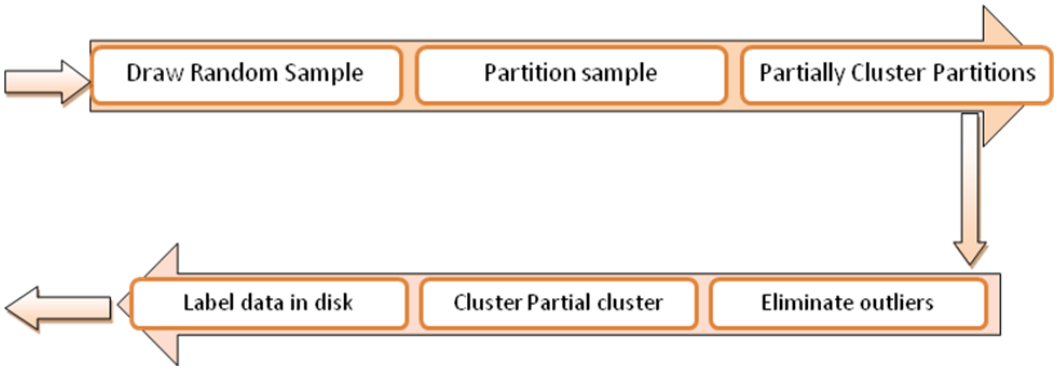


Figure – 5 CURE Algorithm

Advantage of CURE

1. CURE can adjust well to clusters having non-spherical shapes and wide variances in size.
2. CURE can handle large databases efficiently.
3. CURE is robust to outlier.

3.4 Hadoop

It simplifies dealing with Big Data.The Hadoop framework has built-in power and flexibility. The Hadoop framework organizes the data and the computations. The Hadoop solve the Problem of coordinate the work of many computers handle failures, retries, and collect the results together, and so on. The Hadoop Distributed File System (HDFS) provides unlimited file sp-ace available from any Hadoop node. HBase is a high-performance unlimited-size database working on top of Hadoop [11]

Advantage of Hadoop

1. Execution, Efficiency, Scalability, Availability.
2. Solve problems where you have a lot of data perhaps a mixture of complex and structured data.

3.5 Advantages of CURE Agglomerative Hierarchical Algorithm migrate on Hadoop.

1. With non-spherical Shapes of clustering handle large datasets efficiently.
2. It Solve the Problem of Failures, Retries and Collect the results.
3. It reduces the time required for execution.

V. CONCLUSION

In This Paper Cloud Computing is new latest technology available which provides number of services which is useful the client using of internet based application. Cloud Computing provide different types of clustering process which have advantages are available over others. From the analysis of the clustering algorithm hierarchical have more advantages are available. Agglomerative Hierarchical clustering can handle large dataset, increase efficiency of algorithm and parallelism has reduced time required for execution. Every Agglomerative Hierarchical Clustering algorithms have advantage over other. In this select CURE Agglomerative Hierarchical algorithm make a better performance on Hadoop Tool. Improved Performance of Agglomerative Hierarchical Clustering Algorithm is by Hadoop Tool for large set of data. The efficiency of the algorithm is increase. Parallelism of tasks reduces the time required for execution. In future compare CURE Agglomerative Hierarchical algorithm results obtained from cloud platform with m mapreduce framework to understand the effectiveness.

REFERENCES

[1] "The NIST Definition of Cloud Computing"-National Institute of Standards and Technology. Retrieved 24 July 2011.
[2] Data Mining and Cloud Computing by Kushal Venkatesh on 27 November 2012.
[3] Anthony T. Velte, Toby J. Velte, Robert Elsenpter, Cloud Computing (A Practical Approach), McGraw- Hill, ISBN: 978-0-07-162695-8, 2010.
[4] Kriti Srivastava, R. Shah, D. Valia, and H. Swaminarayan," Data Mining Using Hierarchical Agglomerative Clustering Algorithm in Distributed Cloud Computing Environment" published in," International Journal of Computer Theory and Engineering, Vol. 5, No. 3, June 2013".
[5] Esha Sarkar, C.H Sekhar, " Organizing Data in Cloud using Clustering Approach", International Journal of Scientific & Engineering Research, Volume 5, Issue 5, May-2014 , 684, ISSN 2229-5518.
[6] Manpreet Kaur, Mrs.Sukhpreet Kaur02, "Study of Clustering Using Agglomerative and ACO Algorithms", International Journal of Computer Application and Technology (IJCAT) Volume 1Issue 1 (April 2014).
[7] Bhagyashree Ambulkar, Vaishali Borkar, " Data Mining in Cloud Computing", MPGI National Multi Conference 2012 (MPGINMC-2012) 7-8 April, 2012 "Recent Trends in Computing" *Proceedings published by International Journal of Computer Applications® (IJCA)ISSN: 0975 – 8887*.
[8] Hui Gao, *Jun Jiang, Li She, Yan F* , "A New Agglomerative Hierarchical Clustering Algorithm Implementation based on the Map Reduce Framework", International Journal of Digital Content technology and its Applications Volume 4, Number 3, June 2010., doi:10.4156/jdcta.vol4.issue3.9.
[9] B.Thirumala Rao, N.V.Sridevi,V.Krishna Reddy, L.S.S.Reddy, " Performance Issues of Heterogeneous Hadoop Clusters in Cloud Computing", Global Journal of Computer Science and Technology Volume XI Issue VIII May 2011.
[10] *Ms. E. Suganya, Mr. S. Thiruvenkatasamy* , "DISTRIBUTED MINING ALGORITHM USING HADOOP ON LARGE DATA SET",ISSN 2348 – 9928 IJAICT Volume 1, Issue 2, June 2014 Doi:01.0401/ijaict.2014.02.08 Published Online 05 (06) 2014.
[11] Hadoop Beginner's Guide, by Garry Turkington,Packet publishing,Feb-2013.

