Improve Performance of clustering on Cloud Datasets using improved Agglomerative CURE Hierarchical Algorithm

Mrs. Parekh Madhuri H, Prof.Ishan K.Rajani

***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 dendogram. 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.**

***Index Terms*— Cloud Computing, Clustering, Hierarchical Algorithm, Agglomerative Algorithm, Distributed Algorithm, Hadoop.**

I. INTRODUCTION

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.

Manuscript received June, 2015
Mrs Parekh Madhuri H, Computer Engineering, Student, Darshan Institute Engineering & Technology, Rajkot,, Rajkot India,
Prof.Ishan K.Rajani , Computer Engineering, Prof, Dashan Institute Engineering & Technology, Rajkot, India.

Figure-1.1 Cloud Computing Logical Diagram

Advantages of Cloud Computing [2]

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

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
   c. monitoring, logs monitoring and others)
   d. Maintenance (Runtime Governance)
   e. Disaster recovery
6. Lock In.
7. Regulatory and Compliance Restrictions.

II BACKGROUND 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

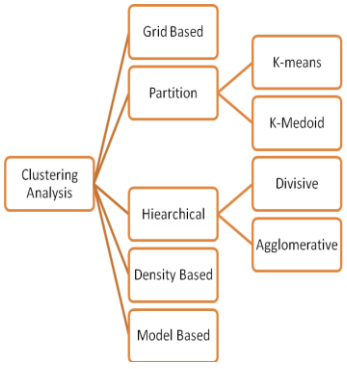

Figure- 2.1 Types of Clustering

ISSN: 2278 – 7798

Agglomerative and Division[4]

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.

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- centroid 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]

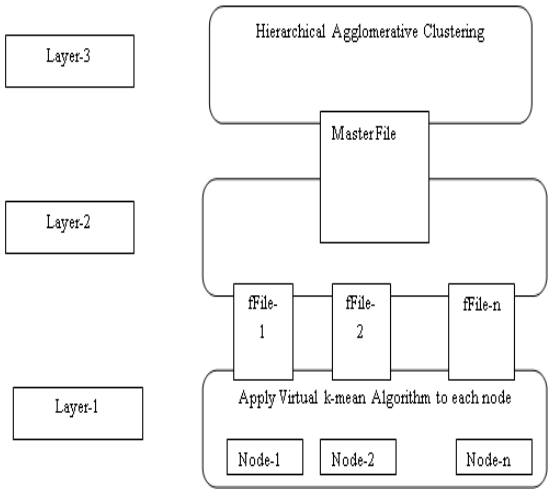

Figure -2.2 Modified Hierarchical Algorithms integrated with k-means

2.3Limitation 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.[4] 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.

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 Improved CURE Agglomerative

Clustering Using Representatives

Improved CURE Hierarchical Algorithm:-[5]

Used Map-Reduce model to decline the size of the database so that retrieval can be done easily and efficiently. Map-Reduce is simple and efficient tool for query processing in a DBMS. For Large data processing task – key advantages of the Map-Reduce framework. [6]

Simple and easy to use: TheMapReduce model is simple but expressive. A programmer defines this task by using only Map and Reduce functions. There is no need for programmer to specify the physical distribution of this work across nodes.

Flexible MapReduce: It does not have any dependency on data model and schema. A programmer can deal with irregular or unstructured data more easily than they do with DBMS.

Independent of the storage: It is independent from underlying storage layers. Thus, MapReduce can work with different storage layers.

Fault tolerance: It is highly fault-tolerant. It is reported that MapReduce can continue to work in spite of an average of 1.2 failures per analysis job at Google.

High scalability: The MapReduce function is of great advantage as it is highly scalable.

Semantic Indexing:- MapReduce provide the easiest way do indexing.

Clustering with Large Dataset:- MapReduce give very easy way for large dataset.

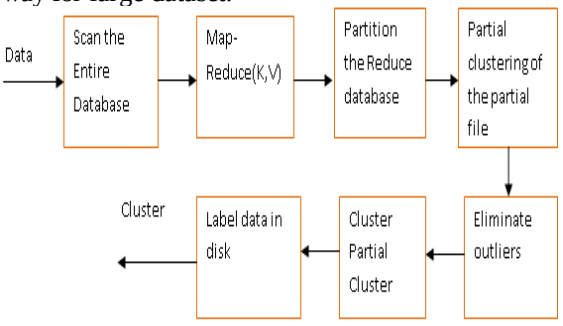

Figure -3.1 Overview of Improved CURE [5]

Step:-1 Scan the Entire Database

Step:-2 collect the reduced data set by using Map-Reduce Technique.

Step:-3 Partition the reduced dataset.

Step:-4 Partitioning the partial file.

Step:-5 Eliminate Outliers.

Step:-6 Cluster Partial Cluster.

Step:-7 Label data in disk.

ISSN: 2278 – 7798

Overview of Map-Reduce Technique [5]

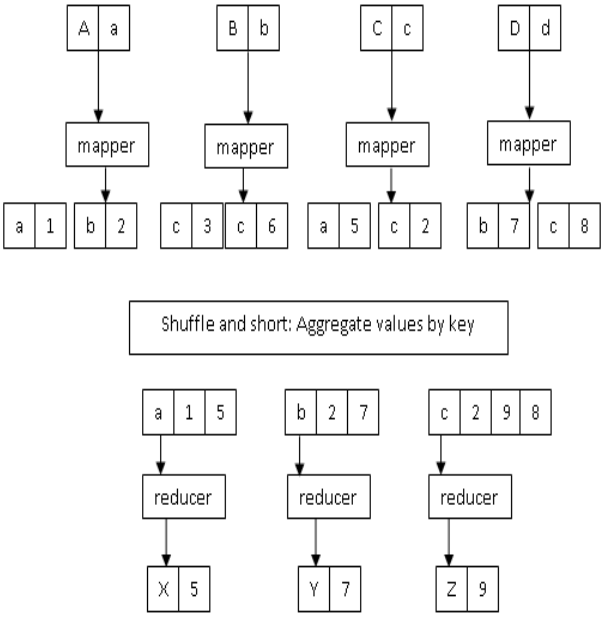

Figure – 3.2 Overview Map-Reduce Technique

Step:-1 Mappers are applied to all input key-value pairs, which are required to generate an arbitrary number of intermediate pairs.

In turn, Reducers are applied to all intermediate values associated with the same intermediate key.[7] There lies a barrier between the map and reduce phase that involves a large distributed sort and group by.

Step:-2 Partitioners and combiners are responsible for different tasks during the clustering of big data.

Partitioners are to divide the intermediate key space, and then to assign the intermediate key-value pairs to reducers.it is responsible for copied. Hash-based partitioner is chosen, and then it starts by computing the hash of the key modulo the number of reducers. Doing this job ensures a roughly even partitioning of the key space.

Among different categories of parctitioners it use hash based partitioners is chosen, and then it starts by computing the hash of the key modulo the number of reducers. Doing this job ensures a roughly even partitioning of the key space.

Mappers ignore values:-

1) Discrepancy in the data handled by reducers.
2) Complex Keys.

Step:-3 Combiners are optimization.

Combiner implemented using local data-structures. The map function only emits once all input records are processed.[8]

Algorithm Improved CURE Hierarchical Clustering. [5]

Data structure used:

a) K: no. of cluster
b) S: size of each cluster
c) Kruskal's MST

Step-1

i) Take an entire database.
ii)Scan the entire database once.
iii) Collect the reduced data set by using Map-Reduce technique.
iv) Determine the total n number of Cluster to be made from the entire database (K).
v) Specify the total number of elements i.e the size of cluster(S).

Step-2

i) Match the data item of a particular characteristic from one cluster with the element in other cluster having the same characteristic.
ii)Search the entire cluster to match.
iii) If match found then merge them.
iv) Update the data structure.

Step-3

i) By the use of Krushkal's MST, the cluster elements with minimum distance are evaluated and finally merged.

Step-4

i) Finally, all the work is collected at one place and each data item with specific characteristic are placed in a particular cluster.
ii)Example: Cluster of Employees, Items etc.
iii) If elements left without match with any item, they will be treated as outliers and kept away from other clusters.
iv) Hence, during clustering, fraud/unwanted/irrelevant data gets trapped.
v) As a result, each cluster has got intra-cluster similarity and inter-cluster dissimilarity.

All the steps used can be depicted as following: [9]

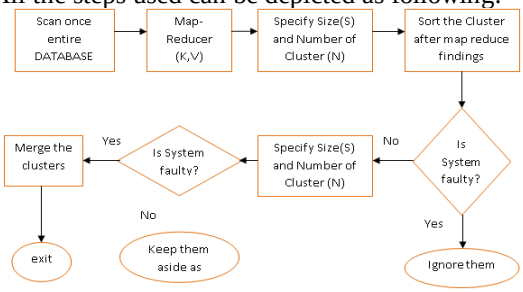

Figure – 3.3 Improved CURE HAC with Map-Reducer

Complexity of CURE Clustering Algorithm. [8]

Time Complexity: = O(n^2)

Space Complexity: = O(n).

Advantage of CURE Clustering Algorithm

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.

Advantage of Hadoop Tool

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

Advantages of CURE Agglomerative Algorithm Migrate on Hadoop

1. It Reduce time required for Execution.
2. With non-spherical shapes of clustering handles large data sets efficiently.
3. It increases efficiency of algorithm.
4. It gives scalability and availability.
5. It solves the problem of Failures, Retries and collects the results.

ISSN: 2278 – 7798

IV. PERFORMANCE

After performing various large data sets on improved k-means and centroid linkage hierarchical algorithm and k-means migrate with improved CURE hierarchical algorithm get the following results. We can conclude here that k-means migrate with improved CURE hierarchical algorithm can prove to be execution time is very less and efficient. And also it handle large dataset with use of hadoop tool and technique is map-reduce.

1. Result Analysis

Instead of centroid linkage hierarchical algorithm used improved CURE hierarchical algorithm and it display how execution time is reduced.

Output

Output of k-means which create cluster?

```
we have got values in a with length : 48
The total time required to run in milliseconds is: 11049.0
```

bin	5/14/2015 2:01 PM	File folder	
build	5/11/2015 7:19 PM	File folder	
centroids_0	5/15/2015 6:13 PM	File folder	
centroids_1	5/15/2015 6:13 PM	File folder	
centroids_2	5/15/2015 6:13 PM	File folder	
centroids_3	5/15/2015 6:13 PM	File folder	
centroids_4	5/15/2015 6:13 PM	File folder	
centroids_5	5/15/2015 6:13 PM	File folder	
src	5/11/2015 7:07 PM	File folder	
.classpath	5/11/2015 7:18 PM	CLASSPATH File	28 KB
.project	5/11/2015 7:05 PM	PROJECT File	1 KB

Figure 4.1 final clusters after k-means

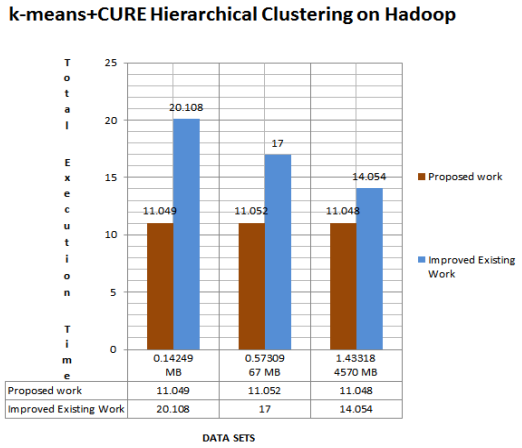

Figure-4.2 Final Result on hadoop with map-reduce with CURE Algorithm

Result as Dendogram of Hierarchical CURE Algorithm

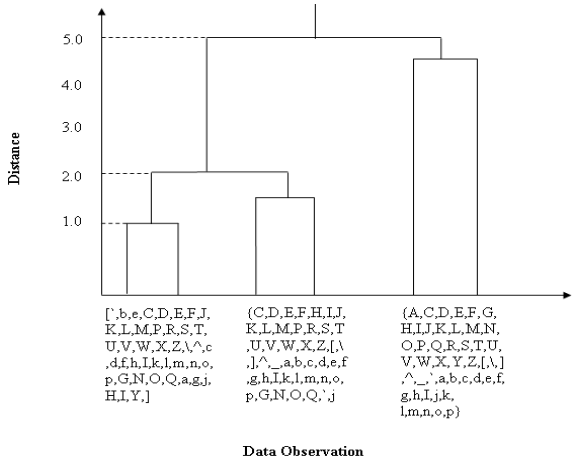

Figure-4.3 Result of Dendogram of CURE Algorithm on hadoop.

Datasets:-USCensus1990.small
http://archive.ics.uci.edu/ml/datasets.html

REFERENCES

[1] 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".

[2] "The NIST Definition of Cloud Computing"-National Institute of Standards and Technology. Retrieved 24 July 2011.

[3] Data Mining and Cloud Computing by Kushal Venkatesh on 27 November 2012.

[4] Anthony T. Velte, Toby J. Velte, Robert Elsenpter, Cloud Computing (A Practical Approach), McGraw-Hill, ISBN: 978-0-07-162695-8, 2010.

[5] An integrated approach for CURE clustering using map-reduce technique Seema Maitrey * , C.K. Jhaa * *Deptt. of Computer Science and Engineering, K.I.E.T., Ghaziabad, U.P., India Deptt. of Computer Science,Banasthali Vidyapith, Newai, Rajasthan.*

[6] K.Lee, Y.Lee, H.Choi, Y.Chung, B.Moon.Parallel data processing with MapReduce: a survey. In ACM SIGMOD, Volume 40 , Issue 4, December 2011 ,Pages 11-20.

[7] Spyros Blanas, Jignesh M. Patel Vuk Ercegovac, JunRao, Eugene J. Shekita, Yuanyuan Tian. A comparison of join algorithms for log processing in MapReduce. Proceedings of the 2010 ACM SIGMOID International Conference on Management of data, Pages 975-986.

[8] M. Thorup. Randomized Sorting in O(nlogn) Time and Linear Space Using Addition, Shift, and Bit-wise Boolean Operations. Journal of Algorithms, Volume 42, Number 2, February 2002, pp. 205-230(26).

[9] Seema Maitrey, C K Jha, Rajat Gupta and Jaiveer Singh. Enhancement of CURE Clustering Technique in Data mining. IJCA Proceedings on Development of Reliable Information Systems, Techniques and Related Issues (DRISTI 2012) DRISTI (1):7-11, April 2012. Published by Foundation of Computer Science, NewYork, USA, and BibTeX.

Mrs. Parekh Madhuri H M.E. (C.E –perusing) B.E.(C.E), Cloud Computing, Distributed Processing.

Prof. Ishan K. Rajani M.E. (C.E.), B.E.(C.E.), Heterogeneous, Parallel and Distributed Computing, working on Coarse and Warp level mechanism of GPUs.

ISSN: 2278 – 7798