Organizing Data in Cloud using Clustering Approach

Esha Sarkar, C.H Sekhar

Abstract--Cloud computing is the latest technology that delivers computing resources as a service such as infrastructure, storage, application development platforms, software etc. Cloud computing is gaining popularity and now-a-days it is on the boom. Huge amount of data is stored in the cloud which needs to be retrieved efficiently. 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. We can integrate data mining and cloud computing (Integrated Data Mining and Cloud Computing– IDMCC) 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 paper, we provide brief description about cloud computing and clustering techniques. Then, it also describes about cloud data mining. This paper proposes a model that applies hierarchical clustering algorithm in the data storage cloud to cluster the data based on the type of data being uploaded by various end users.

Index Terms- cloud computing, cloud storage, clustering, data mining, hierarchical clustering algorithm

1 INTRODUCTION

Cloud computing is getting popular and IT giants such as Goggle, Microsoft, IBM have stated their cloud computing infrastructure. Cloud can be meant as an infrastructure that provides resources /services over the internet. The advantages of cloud computing over traditional computing include agility, lower entry cost, device independency, location independency & scalability. The essential characteristics of cloud computing are: on-demand self service, network access, resource pooling, rapid elasticity, measured service [1]. Cloud computing is a new generation technology that is replacing the other existing technologies as it allows its clients to use its services without worrying about the infrastructure, installation, setup etc and offers them to pay only for what they use. A cloud can be a *storage cloud*, compute cloud or data cloud. A storage cloud provides storage services (block-file based services) that maintains, manages and backs up the enormous data remotely and the users can access it over the network. The main advantage of a storage cloud is that data can be stored virtually. A data cloud provides data management services (record based, column based or object based services) and a compute cloud provides computational services.

Today the cloud architecture is built on top of modern data centers. It incorporates IAAS, PAAS and SAAS. Following figure shows hierarchical view for cloud computing.

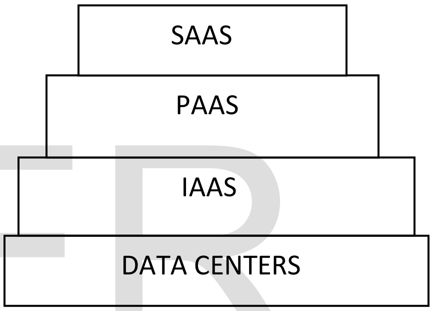

Fig1: Cloud Computing Model

DATA CENTERS: This is the foundation of cloud computing which provides the hardware that the cloud runs on. It is centralized repository for the storage and management of data.

IAAS: Built on top of the data centers layer, IAAS layer provides the computing infrastructure like storage, network connectivity of the data centers servers, firewalls IP addresses etc. e.g.: Amazon EC2, windows azure, Google compute engine.

PAAS: It provides the development platform like operating system, databases, web servers etc. PAAS makes the development, testing and deployment of application quick and cost effective. E.g.: Google app engine, Force.com, Map reduce, windows azure.

SAAS: it provides the software to the end users as a service on demand. SAAS eliminates the need to install setup and run applications on the individual computers. E.g.: Google Map, Google apps, Microsoft 365

- *Esha Sarkar is currently pursuing masters degree program in computer science and engineering in VIIT, India, PH-9704556214 E-mail: esha.sarkar12@gmail.com*
- *C.H Sekhar is currently working as Associate Professor in computer science department of VIIT, India, PH-9949006418. E-mail: sekhar1203@gmail.com*

Data mining is used for extracting potentially useful information from the raw data. Data mining techniques are very much needed in the cloud computing. The implementation of data mining techniques in cloud will allow the users to retrieve meaningful information from non structured or semi structured web data sources [1]. The users need to mine important patterns from the data stored in huge cloud data centers. Thus, the data mining tools play a very important role in cloud computing. The use of contemporary algorithms has proven to be inefficient on the cloud environment. It is not suited for large distributed database as they take very long time for execution. Though there are many algorithms which can handle large database but huge memory usage is the major concern. Using cloud to process and store database can solve this problem as it can take care of more memory requirement very easily [2]. The main objective of this work is to organize the huge heterogeneous data coming from different sources into clusters according to the type of data. This will provide fast retrieval of data from huge cloud data centers.

1.1 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 [2]. It is useful technique for the discovery of data distribution and patterns the underlying data.

1.2 Types of Clustering

The clustering methods can be classified in following:
Partitioning Method: It is a simplest and most fundamental version of cluster analysis which organizes the objects of a set into several exclusive groups or clusters. These are of two types: k-Means and k-Medoids.

Hierarchical Method: It works by grouping data objects into a tree of clusters. The algorithm iteratively split the database into smaller subsets, until some termination condition is satisfied. Hierarchical clustering methods can be classified as agglomerative and divisive. Agglomerative hierarchical clustering is a bottom-up strategy starts by placing each object in 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 are satisfied [4]. Divisive hierarchical clustering is a top-down strategy and does the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster.

Density based Method Clustering is based on density (local cluster criterion). The general idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold. It consists of major features like discover clusters of arbitrary shape, handle noise and one scan. There are mainly three types: DBSCAN, OPTICS and DENCLUE

Grid Based clustering: This approach uses a multi resolution grid data structure. It quantizes the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. STING is a grid based multi resolution clustering technique in which the spatial area is divided into rectangular cells. There are usually several levels of such rectangular cells corresponding to different levels of resolution, and these cells form a hierarchical structure. CLIQUE (Clustering in QUEST) was the first algorithm proposed for dimension-growth subspace clustering in High-dimensional space. In dimension-growth subspace clustering, the clustering process starts at single-dimensional subspaces and grows upward to higher-dimensional ones.

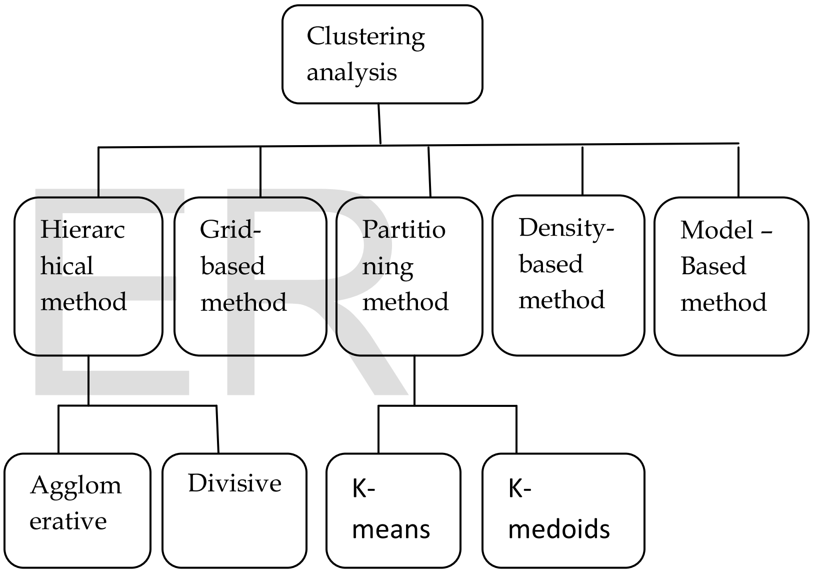

Fig2: clustering methods

Model Based Clustering: It hypothesizes a model for each of the clusters and finds the best fit of the data to the given model. Such methods are often based on the assumption that the data are generated by a mixture of underlying probability distributions. These are of three types-Expectation Maximization, Conceptual clustering and a neural network approach to clustering. The EM (Expectation-Maximization) algorithm is a popular iterative refinement algorithm that can be used for finding the parameter estimates. It can be viewed as an extension of the k-means paradigm, which assigns an object to the cluster with which it is most similar, based on the cluster mean [3].

2 BACKGROUND AND RELATED WORK

One of the cloud services that are being offered is a storage method for the data. Earlier to the concept of cloud computing important industrial data used to be stored internally on the storage media [4]. From music files to pictures to sensitive documents, the cloud invisibly backs up all the files and folders and removes the need for endless and costly search for extra storage space. When there is enormous data, storage cloud alleviates buying an external hard drive or deleting old files to make room for the new ones. Thus many organizations have entered in the cloud environment for the storage service. These organizations pay for the amount of space they use in the cloud. Cloud storage is convenient and cost-effective. It works by storing the files on a server somewhere in the internet rather than on the local hard drive. This allows backing up, sync, and accessing data across multiple devices as long as users have internet capability.

In cloud computing various researches have been made to improve the performance of cloud computing. Various data mining algorithms have been applied in various ways to manage the huge amount of data in cloud. The related works in this field are: Bhupendra panchal and R.K Kapoor [5] proposed clustering and caching methodologies for improving the performance. The main idea is to make replicas of data available at each data centers, so even if one data center goes down, everything in the second data center is clustered with the first. Kashish Ara Shakil and Mansaf Alam [6] proposed an approach that provides management of cloud data through clustering and uses a k-median as clustering technique. A.Mahendiran et al [7] proposed the implementation of k-means clustering algorithm in cloud computing for large datasets. Kriti Srivastava [2] proposed the implementation of agglomerative hierarchical clustering algorithm to enable the benefits such as scalability, elasticity and handling large datasets.

3 PROPOSED MODEL

A storage cloud consists of data coming from different sources and is of different types. The traditional data management techniques are meant for handling traditional data that was of particular type and of limited size, but the data available now is of huge and heterogeneous which can be both structured and unstructured. Thus the traditional techniques fail to handle the requirements of data management in cloud. Both data mining technique and cloud computing helps the business organization to achieve maximized profit and cut costs in different possible ways. One important issue that needs to be dealt with cloud storage is fast accessing of data. Any organization that uses cloud storage expects that it should be able to retrieve information efficiently and in less time. But this expectation is not met due to the fact the data in cloud is unorganized and thus it takes enough time to access the data. There are various end users/ clients of the cloud environment. Each one of the client is using some amount of paid storage space in the cloud to store their huge amount of data.

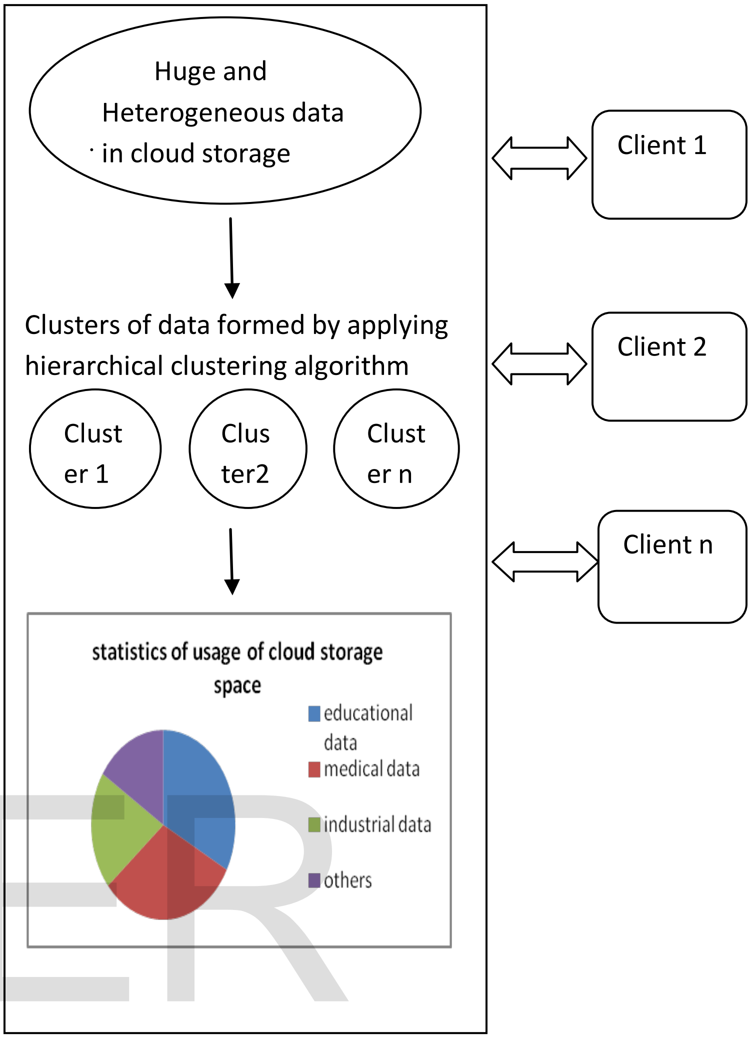

Fig3: Proposed Methodology

Our proposed model is to apply the hierarchical clustering algorithm in the data centers of the cloud storage to form clusters of data based on the type of data being uploaded in the cloud by various organizations. For e.g.: the data from various organizations such as colleges, industries, social networks, hospitals would be stored in the cloud in unorganized manner. Now we apply the hierarchical clustering algorithm in the data centers then the data would be sorted as educational data, medical data, and industrial data and so on.

This will help the end users to retrieve the data quickly. The proposed model will also give the statistics of efficient usage of cloud storage space used by various types of organizations. This statistics will help to manage the data in the data centers and will let the concerned people know about the storage space in the data centers.

4 CONCLUSION AND FUTURE WORK

Cloud storage is a promising technology which is helping large organizations to store and manage their enormous data. Various works has been done in this field to enhance the performance of cloud computing, because one important issue to consider in cloud storage is fast access of data that is stored in the cloud. Our proposed approach provides the implementation of hierarchical clustering algorithm in the cloud data centers to arrange the data according to the kind of data being stored. The proposed method has advantages like 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. Future works for the proposed model is to apply other clustering algorithms in the cloud storage and compare the results to find the best clustering algorithm for cloud storage.

ACKNOWLEDGMENT

I would like to thank and acknowledge my guide Mr. C.H Sekhar for his support, his supervision and his time. His valuable guidance helped me to understand and work in the area of cloud computing.

REFERENCES

[1] Ruxandra Stefania PETRE, "Data Mining in Cloud Computing" published in "Database Systems Journal vol.III, no.3/2012"

[2] 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"

[3] Mrs. Dhanamma Jagli, Mrs. Akanksha Gupta, "Clustering Model for Evaluating SaaS on the Cloud" published in "International Journal of Application or Innovation in Engineering & Management, Volume 2, Issue 12, December 2013"

[4] Naskar Ankita, Mrs. Mishra Monika R, "using cloud computing to provide data mining services" published in international journal of engineering and computer science, volume 2 issue 3 march 2013.

[5] Bhupendra Panchal, R.K Kapoor, "Performance Enhancement of cloud computing with clustering" published in international journal of engineering and advanced technology, volume-2, issue-5, June 2013

[6] Kashish Ara Shakil, Manasaf Alam, "data management in cloud based environment using k median clustering technique" published in "international journal of computer Applications 4th International IT Summit Confluence 2013-The Next Generation Information Technology Summit"

[7] A.Mahendiran, N.Saravanan, N.Venkata Subramanian and Sairam, "Implementation of K-means Clustering in cloud computing environment" published in research journal of applied sciences, engineering and technology.