Clustering in Cloud Computing Environment

Ms.Kiran Dhandore[#1],
Computer Department
Mumbai University
Nerul, Navi-Mumbai, India
kdhandore@gmail.com

Dr. Lata Ragha[#2],
Terna Engineering College
Computer Department
Nerul, Navi-Mumbai, India
lata.ragha@gmail.com

Prof. Sonali Shukla[#3]
Terna Engineering College
Computer Department
Nerul, Navi-Mumbai, India
barveys@gmail.com

***Abstract*—Cloud computing is used to describe a variety of different types of computing concepts that involve a large number of computers connected through a real-time communication network. The large volume of business data can be stored in cloud data centers with low cost. Data mining algorithms are used to extract the useful information from large database. Cluster analysis aims to organize a collection of patterns into clusters based on similarity. Clustering has its root in many fields, such as mathematics, computer science, statistics, biology, and economics. In different application domains, a variety of clustering techniques have been developed, depending on the methods used to represent data, the measures of similarity between data objects, and the techniques for grouping data objects into clusters. The proposed system of Hierarchical Virtual K-Means architecture has developed the hierarchical virtual k-means algorithm to create a net output of homogenous data spread across several machines and data centers in the cloud, which are engaged in storage operations, SAAS, PAAS etc., in different domains of common interest to the users. Also the modified K-means algorithm is proposed for the partitioning and hierarchical environment in the HVKM architecture.**

Keywords— **Data Mining, Clustering, K-mean, Cloud Computing.**

I. INTRODUCTION

Clouds have emerged as a computing infrastructure that enables rapid delivery of computing resources as a utility in a dynamically scalable, virtualized manner. The advantages of cloud computing over traditional computing include: agility, lower entry cost, device independency, location independency, and scalability. Cloud computing is a type of computing that relies on sharing computing resources rather than having local servers or personal devices to handle applications. Recently, a large scale interest has popped up to carry out mining operations in the cloud as the cloud is successfully evolving to accommodate large scale services like SAAS(Software as a Service), PAAS(Platform as a Service). Cloud Computing provides better efficiency and economy in delivering resources on demand.

1.1 Cloud Computing Environment

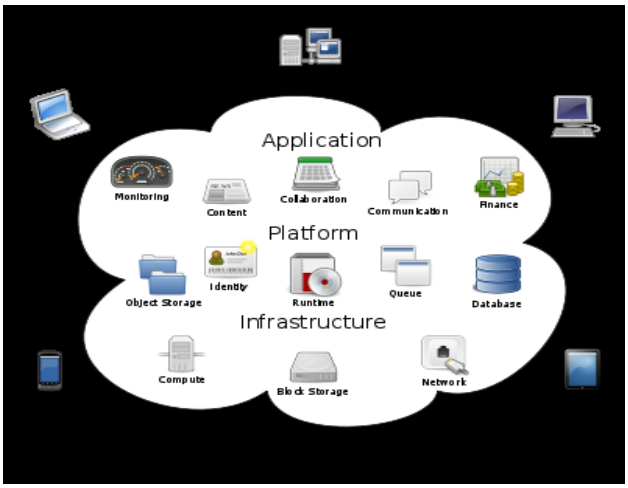

Figure 1: Cloud Computing Environment

Cloud computing is the delivery of computing as a service rather than a product, whereby shared resources, software, and information are provided to computers and other devices as a utility like the electricity grid over a network typically the Internet. A Cloud is a type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and consumers as shown in Figure 1. By distributing and replicating data across servers on demand, resource utilization has been significantly improved. Similarly web server hosts replicate images of relevant customers who requested a certain degree of accessibility across multiple servers and route requests according to traffic load.

Data Centers: This is the foundation of cloud computing which provides the hardware the clouds run on. Data centers are usually built in less populated areas with cheaper energy rate and lower probability of natural disasters. Modern data centers usually consist of thousands of inter-connected servers.

Infrastructure as a Service: Built on top of data centers layer, IaaS layer virtualizes computing power, storage and network connectivity of the data centers, and offers it as provisioned services to consumers. Users can scale up and down these computing resources on demand dynamically. Typically, multiple tenants coexist on the same infrastructure resources. Examples of this layer include Amazon EC2, Microsoft Azure Platform.

Platform as a Service: PaaS, often referred as cloud ware, provides a development platform with a set of services to assist application design, development, testing, deployment, monitoring, hosting on the cloud. It usually requires no software download or installation, and supports geographically distributed teams to work on projects collaboratively. Google App Engine, Microsoft Azure, Amazon Map Reduce/Simple Storage Service are among examples of this layer.

Software as a Service: In SaaS, Software is presented to the end users as services on demand, usually in a browser. It saves the users from the troubles of software deployment and maintenance. The software is often shared by multiple tenants, automatically updated from the clouds, and no additional license needs to be purchased. Features can be requested on demand, and are rolled out more frequently. Because of its service characteristics, SaaS can often be easily integrated with other mash up applications. An example of SaaS is Google Maps, its mashups across from the internet.

1.2 Data Mining Techniques

Data mining is to find knowledge, and knowledge is represented through certain patterns. Association rule is the most often used method in data mining, which finds out the association between data and various objects by finding the potential dependence among data. Classification and clustering can be used to sort out things by characterizing the common significance among different things. There are several major data mining techniques have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns.

1.2.1 Association

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in market basket analysis to identify what products that customers frequently purchase together. Based on this data businesses can have corresponding marketing campaign to sell more products to make more profit.

1.2.2 Classification

Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we make the software that can learn how to classify the data items into groups. For example, we can apply classification in application that "given all past records of employees who left the company, predict which current employees are probably to leave in the future." In this case, we divide the employee's records into two groups that are "leave" and "stay". And then we can ask our data mining software to classify the employees into each group.

1.2.3 Clustering

Clustering is a data mining technique that makes meaningful or useful cluster of objects that have similar characteristic using automatic technique. Different from classification, clustering technique also defines the classes and put objects in them, while in classification objects are assigned into predefined classes. To make the concept clearer, we can take library as an example. In a library, books have a wide range of topics available. The challenge is how to keep those books in a way that readers can take several books in a specific topic without hassle. By using clustering technique, we can keep books that have some kind of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in a topic, he or she would only go to that shelf instead of looking the whole in the whole library.

1.2.4 Prediction

The prediction as it name implied is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables. For instance, prediction analysis technique can be used in sale to predict profit for the future if we consider sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

1.2.5 Sequential Patterns

Sequential patterns analysis in one of data mining technique that seeks to discover similar patterns in data transaction over a business period. The uncover patterns are used for further business analysis to recognize relationships among data.

1.3 Advantages of data mining with Cloud Computing

1. Cloud computing combined with data mining can provide powerful capacities of storage and computing and an excellent resource management.
2. Due to the explosive data growth and amount of computation involved in data mining, an efficient and high-performance computing is very necessary for a successful data mining application.

3. Data mining in the cloud computing environment can be considered as the future of data mining because of the advantages of cloud computing paradigm.
4. Cloud computing provides greater capabilities in data mining and data analytics. The major concern about data mining is that the space required by the operations and item sets is very large.

II LITERATURE SURVEY

Mahendiran et al [2] have implemented K-Means clustering algorithm in cloud computing environment. It is obtained that both Data Mining techniques and Cloud Computing helps the business organizations to achieve maximized profit and cut costs in different possible ways. Thus K-Means clustering algorithm, which is one of the very popular and high performance clustering algorithms, is used in cloud. The main aim of this work was to implement and deploy K-Means algorithm in Google Cloud using Google App Engine with Cloud SQL. Implementation of K-Means algorithm is done in java, so Eclipse IDE is chosen for design and development of the application. To deploy the application in Google, Google App Engine Plug-In is used and Google Cloud SQL is chosen for creating Database and tables.

Michael Shindler et al [3] proposed fast and accurate k-means clustering for large datasets. As cloud computing deals with large data centers, so huge amount of data need to be accessed simultaneously. Thus it impacts on the performance. The fast and accurate clustering algorithm seems to be a better option to adopt in clouds. The results of the work showed that K-Means clustering is much faster than other Divide and Conquer Algorithms.

Ankur Dave et al [4] in Cloud Clustering, toward an iterative data processing pattern on the cloud describe the implementation of Cloud- Clustering, a distributed k-means clustering algorithm on Microsoft's Windows Azure cloud. The k-means algorithm makes a good case study because its characteristics are representative of many iterative data analysis algorithms. They have introduced a general pattern to balance data affinity and fault tolerance for iterative data analysis algorithms in the cloud. They used it to implement k-means clustering and demonstrate its performance. Compared with other patterns, the buddy system is simpler but more robust, as it makes use solely of the reliable cloud messaging service. Furthermore, it can naturally take advantage of the powerful fault domain features provided by Windows Azure, achieving a very efficient balance between data locality and fault tolerance for iterative algorithms.

Clustering is a data mining technique to group the similar data into a cluster and dissimilar data into different clusters. Clustering can be considered the most important unsupervised learning technique so as every other problem of this kind; it deals with finding a structure in a collection of unlabelled data. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression. Cluster analysis is an important human activity. Early in childhood, we learn how to distinguish between cats and dogs, or between animals and plants, by continuously improving subconscious clustering schemes. By automated clustering, we can identify dense and sparse regions in object space and, therefore, discover overall distribution patterns and interesting correlations among data attributes. Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing. In business, clustering can help marketers discover distinct groups in their customer bases and characterize customer groups based on purchasing patterns.

2.1 DISTANCE MEASURE

An important step in any clustering is to select a distance measure, which will determine how the similarity of two elements is calculated. This will influence the shape of the clusters, as some elements may be close to one another according to one distance and further away according to another. For example, in a 2-dimensional space, the distance between the point (x=1, y=0) and the origin (x=0, y=0) is always 1 according to the usual norms, but the distance between the point (x=1, y= [1] and the origin can be 2,$\sqrt{2}$ or 1 if you take respectively the 1-norm, 2-norm or infinity-norm distance.

Common distance functions:

- The Euclidean distance (also called distance as the crow flies or 2-norm distance). A review of cluster analysis in health psychology research found that the most common distance measure in published studies in that research area is the Euclidean distance or the squared Euclidean distance.
- The Manhattan distance (also called taxicab norm or 1-norm)
- The Manhattan distance corrects data for different scales and correlations in the variables.
- The angle between two vectors can be used as a distance measure when clustering high dimensional data.

The Hamming distance (sometimes edit distance) measures the minimum number of substitutions required to change one member into another.

2.2 K-Means Clustering Algorithm

In data mining, *k*-means clustering [6] is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. This results into a partitioning of the data space into Verona cells. K-means (Macqueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain

number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster.

These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point we need to recalculate k new centroids as bar centers of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

2. Assign each object to the group that has the closest centroid.

3. When all objects have been assigned, recalculate the positions of the K centroids.

4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

III HIERARCHICAL VIRTUAL K-MEANS APPROACH

The HVKM system is a data mining approach using partitioning and hierarchical method called Hierarchical Virtual K-Means Approach (HVKM) for the user who desires to provide Business Analysis as a Service. We know where the data of our domain of interest is getting concentrated due to its operation as a unit supporting SAAS (Software as a Service), PAAS (Platform as a Service). Therefore they have integrated such areas for mining purpose through Hierarchical k-means to derive business benefits in domains like POS (point of sales), CRM (customer relation management), international stock market movement, regulatory health patterns etc. For example, in the case of POS and CRM, mining results will help the business analyst to analyse the product movement in the market and the demand for the product.

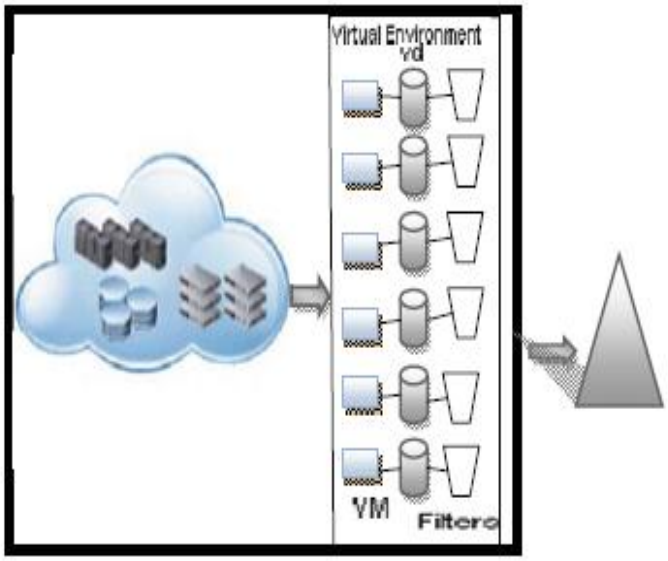

Figure 2: Hierarchical Virtual K-Means Architecture

While approaching the cloud for a data mining operation one of the serious challenges that the operation has to encounter is the successful transactions to be established with the existing virtual machine environment and the databases to be kept under the control. Virtual machine technologies are increasingly being used to improve the manageability of software systems including database systems.

Service provider of cloud computing provides its customers a set of virtual resources such as computation and storing ability [7]. The cloud providers also support a group of native servers and these native servers are further divided consistently to be known as virtual machines. These virtual machines are allotted to the user on request for the cloud. When a request is raised this virtual machines are formed to finish the given task. Hence, when the user is requesting for the data mining in the cloud the native servers will assign the task to the multiple virtual machines where the data will be processed, distributed and managed. When a data mining task is assigned to the virtual machine as mentioned above the virtual resources, virtual development environment and virtual memory are also allocated. Cloud Storage is a model of networked computer data storage where data is stored on multiple virtual servers.

The aim to produce data mining results over the geographically sparse databases under the control of several virtual machines by engaging a single application like HVKM. The multiple databases existing over a cloud which are the domain of interest could also be updated from different applications producing a heterogeneous data update achieved through the flexibility of the middleware architecture prevailing in the cloud. According to the above architecture the data mining operation will be divided into multiple tasks. The mining in multiple servers are done in a hierarchical method. The multiple results of the tasks at different nodes are

hierarchically and algorithmically integrated to form the final results.

The data mining is performed using different servers which are used for data processing in different virtual machines. Clustering is a process of grouping a set of physical or abstract objects into classes of similar objects [4]. Among all the clustering algorithms k-means is a most popular learning algorithm. Hierarchical clustering can be done either with top down or bottom up approach. In top down approach we consider k clusters and divide it into k+1 clusters. In bottom up approach, we merge k cluster into k-1. The goal is to determine the most similar clusters in order to merge the two clusters in to a new one.

HVKM will use the filters to get the temporary result and the process is repeated for the filtered output. The process is repeated in the virtual environment until the desired pattern or output is achieved. In the process of getting the desired output clusters and centroids are formed repeatedly after the data filtering technique is applied to the temporary output, which are called as virtual clusters and virtual centroids. The steps are iterated until a predefined condition is satisfied or the consistency of the virtual clusters does not get affected anymore.

3.1 HVKM Algorithm

The HVKM algorithm is given a predefined number of inputs and the output is decided always based on the requested pattern matching or on the sharpness of the data. The implementation of HVKM is iterative whenever the filtering process is performed. Let *N be* the number of entities, each one is made up of D= {di / i=1…..n} attributes, X={xi/i=1…n} be each data of D. *K* be the number of clusters required, and ***Ci***, $1 \leq i \leq K$, denotes the cluster locations. Each entity is represented in a n dimensional space. Let F be the number of filters done. In HVKM we apply different techniques to find the initial points.

There are few distance metrics such as Euclidian and Manhattan which are used for finding the distances. In order to implement a sample system engaging this algorithm we will consider a Euclidian space metrics.

If **p** = (*p*1, *p*2... *pn*) and **q** = (*q*1, *q*2... *qn*) are two points then the distance from **p** to **q**, or from **q** to **p** is given by
d (p, q) = d (q, p) = **(|p1-q1|2 + |p2-q2|2 + … + |pN-1-qN-1|2 + |pN-qN|2)1**
The method of HVKM implementation can be given in 7 steps. In k-means algorithm initially we define the number of clusters required for the task. In the second step of the k-means algorithm, the distance measurement between each of the sample, within a given cluster, to their respective cluster centroid, is calculated. As the third step, we take the output of the basic algorithm and we filter the output and new clusters are formed basing on the filtered results. The filtering processes in continued until the accuracy of the output data is sharp enough.

The steps for a level 0 node of cloud are as follows:

1. Define the number of nodes N and number of iterations i at level 0.

2. At level 0 and Node 1 virtual machines VM = {vm1…vmn} are allocated.

3. Number of clusters VC = {vci / i=1..n}are formed and k-means algorithm is applied on the clusters.

4. Range method is used to find initial centroids

   VCi = ((pmax–pmin)/k)*n (1)

   VCj = ((qmax – qmin)/k)*n (2)

   Now the centroid is VC (vci, vcj) where p and q values represented in (1) and (2) are the attributes. As mentioned above k is the number of clusters, i , j and n vary from 1 to k where k is an integer. The value (pmax – pmin) will provide the range of p attribute, similarly the value (qmax – qmin) will give the range of q attribute.

5. Find the distance either by Euclidean's distance formulae On the basis of these distances; generate the partition by assigning each sample to the closest cluster.
   Euclidean Distance Formula

$$d(Pi, Pj) = \left(\sum_{k=1}^{m} (Pik - Pjk)^2 \right)^{\frac{1}{2}} \qquad (3)$$

Here d(Pi, Pj) is the distance between P*i* and P*j*. P*i* and P*j* are the attributes of a given object, where i and j vary from 1 to N where m is total number of attributes of a given object. i, j and N are integers. The firsts set of results are generated here. Consider it as VCR1.

6. Item based collaborative filtering algorithm is applied on VCR1, a set of filtered results are obtained at Node 1 i.e. N1 known as VF1.

7. Basing on the Filtering results the new centroid is calculated and the virtual clusters are formed in hierarchical method using bottom up approach. Thus the process is repeated until the desired output is exact in some cases and sharp enough in other cases.

Hence, as the approach of HVKM is hierarchical the level 0 nodes contains clusters VC= {vc1….vcn}, which are considered to be virtual data clusters of the virtual processor VP, where VP = {vp1….vpn} Iteration is done once Vdn data sets are received from Node 0. Filtering processes is done on receiving the data from the first set of virtual processors. The number of filters is equal to the number of virtual clusters formed.

3.3 Modified approach K-Mean Algorithm

The K-mean algorithm is a popular clustering algorithm and has its application in data mining, image segmentation, bioinformatics and many other fields. This algorithm works well with small datasets. Therefore we proposed an algorithm that works well with large datasets. Modified k-mean algorithm avoids getting into locally optimal solution in some degree, and reduces the adoption of cluster -error criterion.

Algorithm: Modified approach (S, k), S={x1, x2… xn}

Input: The number of clusters k1 (k1> k) and a dataset containing n objects(Xij+).

Output: A set of k clusters (Cij) that minimize the Cluster - error criterion

1. Compute the distance between each data point and all other data- points in the set D.
2. Find the closest pair of data points from the set D and form a data-point set Am (1<= p <= k+1) which contains these two data- points, Delete these two data points from the set D
3. Find the data point in D that is closest to the data point set Ap, Add it to Ap and delete it from D
4. Repeat step 4 until the number of data points in Am reaches (n/k)
5. If p<k+1, then p = p+1, find another pair of data points from D between which the distance is the shortest, form another data-point set Ap and delete them from D, Go to step 4

Algorithm A

- For each data-point set Am (1<=p<=k) find the arithmetic mean of the vectors of data points Cp(1<=p<=k) in Ap.
- Select nearest object of each Cp(1<=p<=k) as initial centroid.
- Compute the distance of each data-point di (1<=i<=n) to all the centroids cj (1<=j<=k+1) as d(di, cj)
- For each data-point di, find the closest centroid cj and assign di to cluster j
- Set Cluster Id[i]=j; // j:Id of the closest cluster
- Set Nearest_Dist[i++]= d(di, cj)
- For each cluster j (1<=j<=k), recalculate the centroids
- Repeat

Algorithm B

1. For each data-point di
2. Compute its distance from the centroid of the present nearest cluster.
3. If this distance is less than or equal to the present nearest distance, the data-point stays in the cluster Else ;
4. For every centroid cj (1<=j<=k) compute the distance (di, cj); Endfor

   Assign the data-point di to the cluster with the nearest centroid Cj
5. Set Cluster Id[i] =j

   Set Nearest_Dist[i] = d (di, cj); End for

Modified approach K-mean on the basis of large number of records and execution time using this algorithm. Modified approach K-mean has better performance in comparison to standard K-means algorithm. Modified approach Kmean takes less time of computation time as well as than the K-mean and if the number of clusters is more, then it is again true that Modified approach K-mean takes minimum time to execute than the K-mean.

IV CONCLUSION

Cloud is a highly dispersed computing model which is still evolving, posing multiple challenges for data integration and distribution intended to collect and integrate. In the design of a cloud computing environment where the data sets are available on demand and the basis of the data set we apply further data mining techniques for finding the useful patterns. The mining results also help the management to review the

business strategies and to focus on the customer satisfaction and competitive positioning.

The data mining in Cloud Computing allows organizations to centralize the management of software and data storage, with assurance of efficient, reliable and secure services for their users. Given a data set, the ideal scenario would be to have a given set of criteria to choose a proper clustering algorithm to apply. Most of the algorithms generally assume some implicit structure in the data set. The K-means approach has been used in cloud computing environment and we have proposed a modified K-mean approach which has better performance in comparison to standard K-means algorithm. Modified K-mean approach takes less time of computation than the K-mean and if the number of clusters is more, then it is true that Modified approach K-mean takes minimum time to execute than the K-mean.

ACKNOWLEDGMENT

I would like to express my sincere gratitude towards my guide Dr. Lata Ragha and my co-guide Prof. Sonali Shukla for the help, guidance and encouragement in the development of this methodology. They supported me with scientific guidance, advice and encouragement, and were always helpful and enthusiastic and this inspired me in my work. I have benefitted from numerous discussions with guide and other colleagues.

REFERENCES

[1] N. Janardhan , T. Sree Pravallik , Sowjanya Gorantla "An Efficient Approach For Integrating Data Mining Into Cloud Computing**"** International Journal of Computer Trends and Technology (IJCTT) - volume4 Issue5–May 2013, pp 1291-1294 ISSN: 2231-2803

[2] A. Mahendiran, N. Saravanan, N. Venkata Subramanian and N. Sairam "Implementation of K-Means Clustering in Cloud Computing Environment" Research Journal of Applied Sciences, Engineering and Technology 4(10): 1391-1394, 2012 ISSN: 2040-7467.

[3] Michael Shindler, Alex Wong and Adam Meyerson "Fast and Accurate k-means For Large Datasets", 25th Annual Conference on Neural Information Processing Systems (NIPS) 2011, pp.1-9

[4] Ankur Dave" Cloud Clustering: Toward an iterative data processing pattern on the cloud" 2011 IEEE International Parallel & Distributed Processing Symposium, pp. 1132-1137

[5] Jiawei Han, Micheline Kamber." Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, Champaign CS497JH, fall 2001,http://www.cs.uiuc.edu/hanj/bk2

[6] G. Malathy and Rm. Somasundaram, "Performance Enhancement in Cloud Computing using Reservation Cluster", European Journal of Scientific Research, ISSN 1450-216X, Vol. 86 No 3, September, 2012, pp.394-401

[7] Bhupendra Panchal and R. K. Kapoor, "Performance Enhancement of Cloud Computing with Clustering", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-5, June 2013, pp.189-191

[8] T.R.Gopalakrishnan Nair, K. Lakshmi Madhuri," Data Mining Using Hierarchical Virtual K-Means Approach Integrating Data Fragments in Cloud Computing Environment " Proceedings of IEEE CCIS2011 pp. 230-234

[9] Wei-Tek Tsai, Xim Sun, Janaka Bala Sooriya. "Service Oriented Cloud Computing Architecture" IEEE 2010 Seventh International Conference on Information Technology, pp 684-689