ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

Survey on Cloud Storage Based Clustering Technique

P.Suresh,R.Malathi

Research scholar, Dept. of Computer Science, H.H The Rajahs College (Autonomous), India

Assistant professor, Dept. of Computer Science, H.H The Rajahs College (Autonomous), India

ABSTRACT : Cloud storage enables users to remotely store their data and enjoy the demand high quality cloud application. Without burden of local hardware and software management through the benefits one clear such a service is also relinquishing users physical possession of their outsourced data, with inevitably poses new security risk towards the correctness of the data in cloud. cloud storage service avoid the cost expensive on software, personal maintains and provides better performance less storage cost and scalability, cloud service through internet which increase their exposure to storage security vulnerabilities however security is one the major drawbacks that preventing large organization to enter into cloud computing environment. This work surveyed on several storage techniques with using clustering techniques.

KEY TERMS: cloud; cloud storage; cluster; data mining; distance measure.

I.INTRODUCTION

Cloud computing is getting popular and it giants such as GOOGLE, MICROSOFT, IBM have stated their cloud computing infrastructure. Cloud can be meant as an infrastructure that provides resources/services over the internet. The advantage of cloud computing over traditional computing include agility, lower entry cost, devices independency, location indecency & scalability. cloud computing is a new generation technology that is replacing the other existing technology as it allows its client to use its service without worrying about the infrastructure, installation, setup etc and offer them to pay only for what they use. A cloud can be STORAGE cloud, COMPUTE cloud, DATA cloud. A STORAGE CLOUD provides storage services that maintains, manages and back up the enormous data remotely and the users can access it over the network. cloud computing offer served benefits to the business organization to cut the initial investments to establish infrastructure for storage and compute many business organization have already stored migration of their business data into mine useful information from the data stored in the cloud data center with regards to business decisions. main objective of this work is to incorporate and implement K-MEANS data mining technique into cloud environment .K-MEANS clustering algorithm is one of very popular and high performance clustering algorithm each point is assigned to a cluster with the closest centroid, the main of cluster must be known in advance, which is defined by k.

*select k-means as the initial centroids.
*Repeat
*From k-cluster by assigning all points to close centroid.
*Recompute the centroid of each cluster.
*untill the centroids don't change.

The initial certroids will be chosen randomly. the centroid is nothing but mean of the points in the in cluster. Euclidean Distance measure the closeness. k-means generates different cluster in different runs.

II.LITERATURE SURVEY

Cloud storage service of permits consumers to the data in cloud as well as allowed to utilize the available well qualified application with no worry data storage maintains the Flexibility and scalability ,the characteristic of flexibility /elasticity

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

which provisions resource according to demand to scale horizontally by increasing number of system and vertically by increasing and decreasing hardware configurations .cloud computing abstract the physical resource ,results in, user have sense of location independence . Data mining is the process of finding correlation or patterns among dozens of field in large relational databases. Many companies follow data mining in order get the consumer focus. Data mining is used in many field such as retail, medical, organization, etc. they are method associated in knowledge extraction. Clustering is the one of the popular method in data mining, it the dataset based on some similarities. in definition it can be described as "it is the task of grouping a set of objects in the same group(cluster) are more similar to each other than to those in other group (cluster).in partitioned clustering, the objects in dataset are relocated by moving for one cluster to other based on some computational value. It is also known as the method, the number of cluster should be predefined by the user. Namely a relocating method iteratively relocates points between the k-cluster. K-mean algorithm is the most popular algorithm in partitioned clustering.

III. **ABOUT CLUSTER ANALYSIS**

The process of grouping a set of physical or abstract objects into classes of similar object is called clustering. a cluster is a collection of data object that are similar to one another with in same cluster and are dissimilar to the object in other cluster. clustering is also called data segmentation in some application because clustering partitions large data sets into groups according to their similarity. clustering in data mining scalability, ability to deal with different type of attributes ,discovery of cluster with arbitrary shape, high dimensionality, constraint based clustering. the clustering methods can be classified in following servel type partitioning method is a simplest and most fundamental version of cluster analysis is partitioning which organizes the objects of a set into several exclusive groups or cluster. these are of two k-means and k-medoids .hierarchical method works by grouping data into a tree of cluster. divisive hierarchical clustering is a top-down strategy does the reverse of agglomerative hierarchical clustering by starting with all object in one cluster. density based method: clustering based on density (local cluster criterion).such as density connected points. it consists of major feature like discover cluster of arbitrary shape, handle noise and scan. they are three types DBSCAN,OPTICS and DENCLUE

.

DBSCAN (Density Based Spatial Cluster Of Application With Noise) is a density based clustering algorithm. the algorithm grows regions with sufficiently high density into cluster and discovers clusters of arbitrary shape in spatial databases with noise.it define a cluster as a maximal set of density connected points.

OPTICS: ordering points to identify the database with respect to its density based clustering structure.this cluster ordering contains information equivalent to the density based clustering corresponding to a board range of parameter settings.good for both automatic and interactive cluster analysis, including finding intrinsic cluster structure.

DENCLUE:(DENsity-based CLUstEring)is a clustering method based on set of density density distribution functions. grid based clustering approach usesa multi resolution grid data structure. it quantizes the object the space into a finite number of cells that from a grid structure on which all of the operations for cluster are performed. STING is a grid based multi resolution clustering technique in which the spatial area is divide into rectangular cells.

CLIQUE:(CLustering in QUEst) was the fast algorithm proposed for dimension growth subspace clustering in high dimensional space.in dimensional growth supspace clustering, the clustering process start at single-dimensional subspaces and grows upward to higher dimensional space.

MODEL BASED CLUSTERING: method attempts to optimize the fit between the given data and some mathematical model.such method are often based on the assumption that the data are generated by a mixture of underlying probailty distributions.these are of three types excection maximization ,conceptual clustering and a neural network approach to clustering.the em(expectation maximization) algorithm is a popular iterative refinement algorithm that can used for finding the parameter estmates.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

IV.TIME COMPLEXITY OF CLUSTER

The k-means algorithm converges to local minimum. Before the k-means converges, the centroids computed number of times, and all points are assigned to their nearest centroids. Where p is the number of points, k is the number of cluster and t is the number of iterations. In this proposed enhanced k-means clustering algorithm, to obtain initial clusters, this process requires $0(pkt)$,.here some points remain in its cluster, thers move to another cluster. if the points stays in its cluster this require $O(1)$.

V.ABOUT K-MEANS

The k-means algorithm is an iterative procedure for clustering which require an initial classification of data. it computes the center of each cluster, and then computes new partitions by assigning every object to the cluster whose center is the closet to that object. this cycle is repeated during given number of number of iterations or until the assignment has not changed during one iteration. this algorithm is based on an approach where a random set of cluster base is selected from the original dataset, each element update the nearest element of the base with the average of its attributes. The k-means is possibly the most commonly used clustering algorithm. It is most effective for relatively smaller data sets. The k-means finds a locally optimal solution by minimizing a distance measure between each data and its nearest cluster center. The basic k-means algorithm is commonly measured by any of intra-cluster or inter-cluster criterion.

K MEANS STEPS

```
INPUT:D=DATASET                                  11.end if
K=THE Number of centers                          12.end for
OUTPUT: Set of k centroid cϵ c repreting a good partitioning of D into k clusters
                                                 13.if di, center then
1 select the intial cluster centroids c          14.changed ++
2 repeat                                                 15.Recompute cj new for
3  changed=0 \\find the closest toevery data point d...
4 for all data point di ϵ  D do                  16.end if
5 assigned center=di.center                      17.end for
6 for all center cj ϵ c do                       18.until changed==0
7 compute the squared Euclidean distance dist=dist (di,dj)
8 if dist<di center distance then
9 di center distance=dist
10 di center=j
```

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

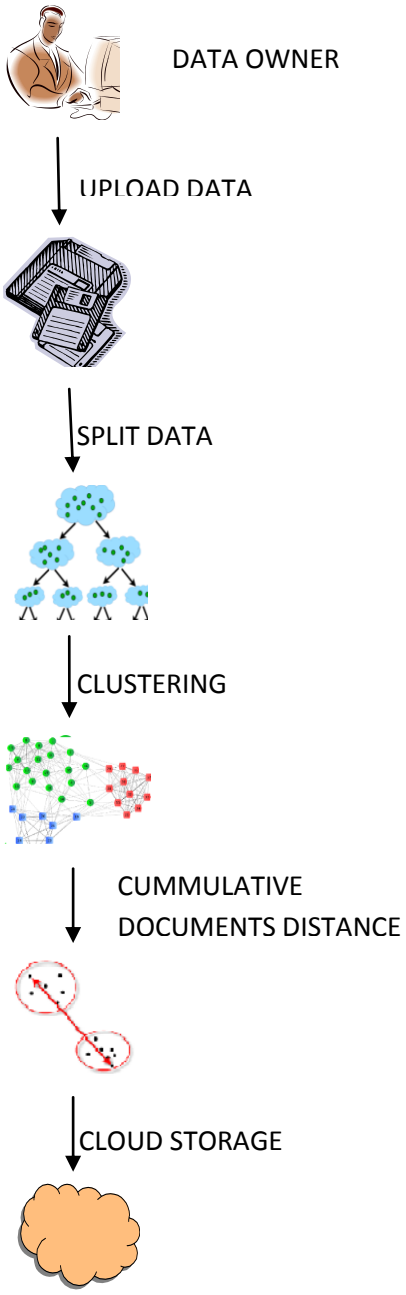

Figure:1 cloud storage file in clustering technique

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

K-MEANS APPLICATIONS

Are different type of method to using data mining. They are several type of method ,optical character recognition, biometrics, diagnostic system and military applications.

TIME COMPLEXITY K-MEANS

Computing distance between doc and cluster is 0(m) where m is dimensionality of vectors. Reassigning cluster:0(kn)distance compution,0(knm).computing centroid each doc gents added once to some centroid 0(nm).assume these two steps are each done once for iterations 0(1knm).

PERFORMANCE MATRICES OF K-MEANS

It is recommended to run k-means several times to achieve the correct result.To avoid the problem described above, the cluster center are initialized with randomly chosen data points.If D_{ik} becomes zero for some X_k ,singularity occurs in the algorithm, so the initializing center are not exactly the random data points, they are just near them.(with a distance of 10^{-10} in each dimension)

if the initialization problem still occurs for some reason(e.g the user adds wrong initialization of the function),the "lonely" centers are redefined to data points.

VI.K-MEANS ALGORITHM

For corresponding give the data set x, choose the number of cluster $1 < c < N$. initialize with random cluster center chosen from the data set.

REPEAT for$l = 1,2 \ldots \ldots$

STEPS 1 compute the distances

$$D_{IK}^{2} = (Xk - vi)^{T}(Xk - vi), 1 \le i \le c, 1 \le k \le N$$

STEP 2 select the points fro a cluster with the minimal distances, they belong to that cluster.

STEP 3 Calculate cluster center

$$v_i^{(l)} = \frac{\Sigma_{j=1}^{ni} xi}{ni}$$

until

$$\Pi_{k=1}^{n} max \left| v^{(1)} - v^{(l-1)} \right| \neq 0.$$

Ending Calculate the partition matrix.

NETWORK MODEL K-MEANS

Consider a sensor network consisting of a large number of light weight, wireless, battery-power sensors for environment monitoring. each sensor is for environment monitoring. each sensor is measuring the same variables and clustering all the data in the network in a given time window can offer variables and clustering all the data in the network in a given time widow can offer valuable information concern environmental phenomena. since the power required for wireless

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

communication goes up with the square of distance, nodesonly should communcate with others in a small radius(immediate neighbors).moreover, the large number of nodes makes global syncronization undesirable. thus , a k-means clustering algorithm in this example ought to workin a locally synchronous manner i,e nodes only synchronize with their immediate neighbors.

VII.K-MEANS PERFORMANCE FILE

Number cluster (k)	Performance			
	SOM	K-MEANS	EM	HCA
8	59	63	62	65
16	67	71	69	74
32	78	84	84	87
64	85	89	89	92

Table 1: No of cluster and the performance on different algorithm.

According to the number of cluster. k(see table 1) expect for hierarchical clustering, all clustering algorithm compared here require setting k in advance (for SOM, k is the number of nodes in the lattice).here, the performance of different algorithm for different k, is compared i order to test the performances that are related to k. to simplify the situation and to make the comparison easier .k is chosen equal to 8,16,32 and 64 and the lattices for square them.

PERFORMANCE INDEX	
70 AND ABOVE	EXCELLENT
60-69	VERYGOOD
50-59	GOOD
45-49	VERY FAIR
40-45	FAIR
BELOW	POOR

Table 2::Performance index on k-means performance file.

VIII.ABOUT EUCLIDEAN DISTANCES MEASURE AND STEPS

A number of different measures have been proposed to measure 'distance' for binary and categorical data. for details see the book by everitt, landau and leese. readers are also referred to this text for details of what to do if you have a mixture of different data types. for interval data the most common distance measure used is the Euclidean distance.

Euclidean distance in general, if you have p variables $x_1, x_2 \ldots, x_p$ measured on a sample of n subject, the observed data for subject i can denoted $x_{i1}, x_{i2}, \ldots, x_{ip}$ and the observed data for subject j by $x_{i1}, x_{i2}, .., x_{ip}$. the euclidean

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

distance between two subjects is given by $d = \sqrt{(x_{i1} + x_{j2}) + \left(x_{i2} + x_{j2}\right) \ldots + \left(x_{ip} + x_{jp}\right)^2}$ when using a measure such as the euclidean distance,the scale of measurement of the variables under consideration is an changing the scale will obviously effect the distance between subject (e.g a different of 10cm could being a difference of 100mm).in addition, if one variable has a much wider range than this variable will tend to dominate. for example, if body measurements had been taken for a number of different people, the range (in mm)of height would be much wider than the range in wrist circumference, say. to get around this problem it tends to reduce the variability (distance) between clusters. this happens because if a particular variable separates observation well then, by definition, it will have a large variance (as the between cluster will become less. despite this problem many textbook do recommend standardizing and once with to see how much difference, if any, this make to the resulting clusters.

ABOUT CLOUD STORAGE

Our lives have a huge relationship with computer and we need use computer everywhere in our daily routine. nowadays, no-one live without computer and the computer use everywhere, such as, governments, companies, school, university even home has at one computer. everyone who uses the computer has his different varied of files, and the problem is they need to use files in different place even in their Smartphone. cloud storage makes their life easier by saving their files in the internet and they have an access to opening, copying, editing, and adding by using any computer or smart phone any where."with the rise of wireless mobile devices, such as cellular phones and net books, they is an increasing need to access internet resources and move private data between devices. Using cloud storage as the medium for such information exchange is attractive. Therefore, shall talk extensively about what is cloud storage and how can it be useful. This will also tackle several of aspects that concern cloud storage such as social, legal, ethical and security concerns.

BACKGROUND CLOUD STORAGE

The cloud storage is a server in the internet has hundreds and thousands of hard drives to save their customers files or their employee files for sharing and open them any place has internet. before the cloud storage existed, people was use vary type of storage. let us go back years ago, people were able to save their files in a floppy disk and that could carry only 1.4 megabytes(MB) and "small computer often use floppy disk for storage"(pechura,1983).in that time, it was enough to carry a small program or a couple of files. after that, cd become up and that was a high discover for how cared about storage because it could carry 700mb of data. a few years later, computing people discovered DVD and this was a huge new technology which can carry 4.7 gigabytes(gb) to 17.08 in a different capacities, while saving files increasing and has different and has different types of storage, people somehow need to have one place for their files and they found it in the cloud storage. Today, cloud storage can about this file to save them for a long time.

BENEFITS OF CLOUD STORAGE

Cloud storage has money benefits for the users. first of all the users can save his files forever in the internet forever if the planned to use a small space and there are many companies offer a small space for free, such as, google, hotmail, and dorpbox. for example, Google drives the user 5GB for free which allows user to save his files in Google server.

IX.CONCLUSION

In this paper, first prop osed a new architecture for cloud data storage in which the private cloud should store only within organization of sensitive data. This is to provide a simple and qualitative and power algorithm and the Euclidean distance as a measure similarity distance. To demonstrated our technique using k-means clustering algorithm. K-means found to be in the top data mining algorithm were identified by the ieee international conference on data mining. Despise its draw backs, k-means remains the most widely used partitioned clustering algorithm in practice. The algorithm is simple, easily understandable and reasonable, scalable, and can easily modify to deal with streaming data in sensible grouping which

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

arises naturally in various scientific fields.Because of this reason it is not surprising to notice the increasing vogue data clustering it is significant to remind that cluster analysis is an exploratory toll, output of clustering algorithm only suggest hypotheses. On the contrary, thousand of enhanced clustering algorithm have been seen and new area continue came into exist. To find the best algorithm is again the part of research but among research developments, most algorithms. The k-means using Euclidean distance measure technique successfully implementation. K-means method in efficient way, we are looking to proposed using technique Euclidean distance measure to implement centralized user with organization.

REFERENCES

[1] K.Alsabti,S.Ranka,And V.Singh.An Efficient K-Means Clustering Algorithm.Http:\\Cise.Ufl.Eduranka\.1997.
[2] Comparison Between Data Cluster Algorithm ."The International Arap Journal Of Information Technology .Vol 15,No 3, July 2008.
[3] R.C Dubes And A.K.Algorithm For Clustering Data.Prentic Hall.1988
[4] L.K Kaufman And P.J.R Rousseeuw.Finding Groups In Data ,An Introduction To Cluster Analysis .John Wily Sons,1990.
[5] C.Wang,Q.Wang.K Ren,And W,Lou"Ensuring Data Storage Security In Cloud Computing."Pro Of 'W Qos '09;2009 Julys.
[6] Wang J, Wan,Liv,Z And Wang P.2010'Data Mining Of Mass Storage Based On Cloud Computing".In Processing Of 2010 Ninth International Conference On Cloud Computing.
[7] The Text Book "Data Mining Concept And Techniques" By Jiawei Han, Michline Kamber,3rd Edition.
[8] Fahim A.M Salem A.M,Tortly F.A Ramandan M.A "An Efficient Enhanced K-Means Clustering Algorithm.1626 A.L./Jzheejiang Univ Science A 2006.
[9] Jain A,Murthy.And Flynn P,"Data Mining :A Survay"ACM Computing Survay,Vol 31.No3.1999.
[10] Garbriele Derban And Grigoreta Sofia Moldvan,"A Compression Of Clustering Techniques In Aspect Mining",Studia University'vol Number 1,2006;
[11] Shina,Liu Xumin,Gvan Yong "Research On K-Means Clustering Algorithm "3rd Inter-National Symposium On Intelligent Information Technology And Security Informations.
[12] Mps,Bhatia &Deepika Khurana "Analysis Of Initial Confers Fork-Means Clustering Algorithm"Ijca Volume 71,No5 Pp9- 13.May 2013.
[13] B.Duran And P.Odell.Cluster Analysis:Analysis.A Survay.New York:Springer-Verlag .1974.
[14] B.Everitt.S. Landv.And .M.Leese.Cluster Analysis London Arnold 2001.
[15] Garbriela Derban And Grisoreta Sofia Moldvan,"A Comparison Of Cluster Technique In Aspect Mining",Studia ,University,Vol Li Number 2006.Pp 09-78.
[16] Naha Aggarwal, Kirti Aggarwal,Kanika Gupta "Comparative Analysis Of K-Means Cluster Algorithm For Data Mining Ijser August 2013.
[17] Paulbradly And Sama Fayyad:Refining Initial Points For K-Means Clustering,5th International Conference On Machine Leadring 1998.
[18] D.Feely And A,Noore:Excending K-Means With Efficient Estmation Of The Number Of Cluster:In Proceeding Of The Seventeeth International Conference Machine Learning ,San Frocesco,2000.
[19] Perklis Andritosos Data Clustering Techniques, Department Of Computer Science March 11,2002.
[20] Kioti Aggarwal,Neha Aggarwal,Sunita Bhardwaj,Nikita Taneja,"An Effective Enhance K-Means Clustering ,Data Mining ,International Conference On Emerging Trends In Engineering And Management "In Press.
[21] Kumar S.P ,Subramaniyan R(2011)"For Efficient And Secure Storage Security Cloud Computing (Ijocs)Vol18.
[22] Sajthabanu S,Raj E.G (2011).Data Storage Security In Cloud International Journal Of Computer Technology.
[23] Dochmukh P.M ,Gughane As Et Al (2012)"Maintaining Files Storage In Cloud Computing (Ijetae) 2009.
[24] Wang ,W,Li Zental (2009) Secure And Efficient Access To Outsource Data,Ccsw 109'proceeding Of The 209.Acm Worshop Cloud Storage Security.
[25] Ensuring Data Storage Security In Cloud,Iosr Journal Of Engineering.
[26]Zhijun,Wang And Zhang Ni "A Survay On Cloud Computing Security".
[27] Kulkarni,Gurudett,Jayant Gambhir,Tejswini Patil,And Amoula Dongare" A Security Aspects In Cloud Computing"In Softwae Aspects In Cloud Computing "(Icsess)2012.
[28] Yandong,Zhang Yongsheng "Cloud Computing Nand Cloud Security Challenges In Information (2012)Ieee,.
[29] Cong Wang,Qian Wang ,Kuiren,Ning Eao And Wejing Lou,11 "Toware Secure And Dependable Storage Services In Cloud Computing"Ieee 2012.
[30] Mehadi Hojabr, Ensuring Data Storage In Cloud Computing With Effect Of Kerberes"("Ijert".,2012).

BIOGRAPHY

R.Malathi is working as an Assistant professor in the Department of Computer Science , H.H Rajahs College (Autonomous), in Pudukkottai, from tamilnadu. India

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

P.SURESH is a Research Scholar, Department of Computer science, H.H Rajah's College (Autonomous) in pudukkottai, from tamilnadu. India