Cluster Analysis for Cloud Data through K-Means Technique

[1]Banshidhar Chaudhary and [2]Prof. VipinSaxena
[1]banshidharphd@gmail.com, [2]profvipinsaxena@gmail.com
Department of Computer Science BabasahebBhimraoAmbedkar University, Lucknow, India

ABSTRACT

In the current scenario, many of organisations are transferring large databases over the cloud and for which public/private clouds are available. The sizes of the database are increasing in an exponential manner and due to heavy file size of video files, it is necessary to store large database over the cloud. In the present work, a concept of data mining is used to create the different kinds of clusters and further K-means method is applied for analysis of clusters to find out the duplicate pattern in the database, which may in the forms of duplicate files or duplicate folders. A sample database of credit cards is considered for analysis purpose and computed results are obtained in the form of tables and graphs. Validations are also performed by performing the various feasible queries. After setting the duplicate pattern in database, the same are stored inside the clusters and removed by performing feasible queries.

Keywords: Sample Database, Cloud Data, Hidden Pattern, Cluster Analysis, and K-mean Technique
DOI Number: 10.14704/NQ.2022.20.16.NQ88073 **NeuroQuantology2022; 20(16):725-737**

1. INTRODUCTION

Due to evolution of e-commerce, there is tremendous growth of business organisations, which are working 24*7. Many of small organisations have also been merged into large organisations for example in India, Dena bank has been merged into the Bank of Baroda; Oriental Bank has been into Punjab National Bank, etc. Sale and purchase by the use of web pages have been increased over the internet and this was due to Covid-19, which has been spreaded across the globe since 2019. Hence, there is huge increase in the systematic development and hosting of webpages over the internet. All of the websites are handling the digital transactions by the use of credit and debit cards.

A lots of work has been done on the structural database, later on object-oriented databases and due to tremendous growth of Graphical User Interface (GUI), many of organisations are using the concept of distributed database over the cloud. The following figure 1 represents storage of large database over the cloud called as cloud storage which has high end finite number of servers are arranged by the use of topological structure which may be static or dynamic, some of the organisations have own servers for providing the services to the customers

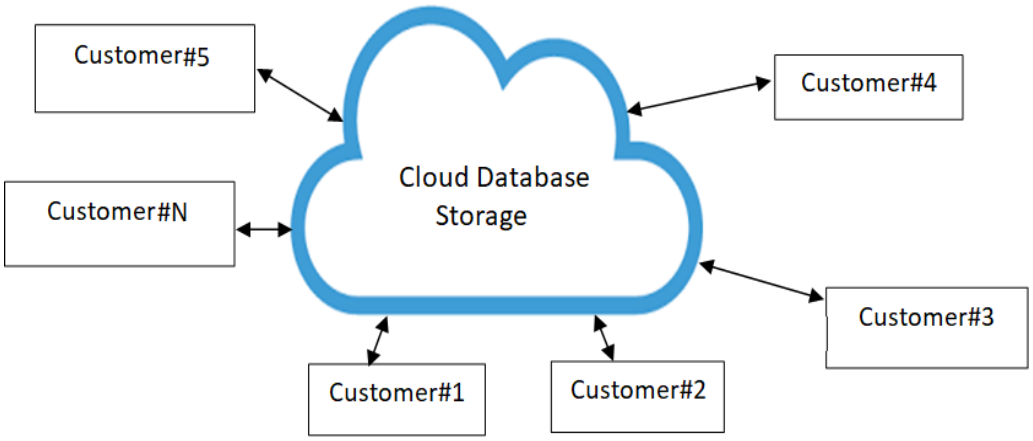


Figure 1. Customer Interaction across Cloud Database

In the above figure, finite numbers of customer's works as cloud users who are accessing the cloud storage. The customers may be more than two and all the customers may access the cloud database storage through one to one, one to many, many to one and many to many communication techniques which further follows the peer to peer or broadcast communications. The cloud storage contains huge amount of database related to credit/debit cards of Indian, overseas banks and duplicacy may also occur at the end of cloud storage, and therefore design of clusters over the cloud is an approach to segments the duplicate database and later on it can be discarded from the cloud storage.

In the present work the concept of cloud database storage, which is attached with finite number of customers, a technique of data mining is used for large and complex database of credit/ debit cards of the bank. The data is separated over the cloud by the use of cluster techniques in which different patterns of clusters are defined and by the use of k-means technique duplicate clusters are identified and later on removed through various queries and thereafter customers are able to access the unique cloud database storage.

2. RELATED WORK

Although lots of research work has been done on the structural and object-oriented database but limited research, work is available for the cloud storage of database; let us describe some of the important research papers related to present work. Matheus.C. J et al. [1] studied and designed a system that is based on a comparison of the system such as Knowledge Discovery Workbench (KDW) given. Kevin Zhu et al. [2] studied the theoretical framework of e- business values in the financial service industries. Yu-Ju-Lin et al. [3] used a neural network algorithm for building a decision model for insurance with characteristics like the health and wealth of an investor. Shu-Hsien Liao et al. [4] investigated the best functionality using the Apriori algorithm and clustering analysis of data mining for the life insurance product. Hurtado, C et al. [5] presented a concept based clustered approach to finding proteins and genes that have similar functionality. Jayanti R. et al. [6] proposed an algorithm and highlighted the data mining technique for Customer Relation Management (CRM) in a real case of the insurance industry to find hidden and unknown information. Shu-Hsien Liao et al.
[7] have presented an analysis of the clustering of consumer preferences in hypermarkets. Fariba Sadri [8] surveyed Ambient Intelligence (AmI); a true range of applications included in this survey presented a challenge of ethical, social and technical. Sharma. R. the et al. [9] have implemented k–mean clustering analysis of crop yield using the Weka interface. Shanshan Li et al. [10] have presented the performance of both neighboured-constrained clustering algorithm and advanced Spectral – Spatial-clusteringalgorithm. Mauricio C et al.

[11] surveyed of reviewed the discovery of domain-specific query forms.

Mousazadeh. S and Cohen. I [12] developed a Voice Activity Detection (VAD) algorithm based on the spectral clustering method in the presence of transient noise. Djidjev et al. [13] showed the problem of finding a partition maximization of the modularity of a given graph, which can be reduced to a minimized weighted cut problem on a complete graph. Papalexakis. Iet al. [14] demonstrated a need-based co-clustering on multilinear decomposition with the Sparse Latent factor. Ali. M et al. [15] proposed an innovative probabilistic clustering concept for wind form aggregated models. Joseph. R et al. [16] proposed a decision-making method based on scenario discovery, which uses a statistical data-miningalgorithm. Kumar, N et al. [17] have to cluster a data set, which is insurance data with the use of k-mean.

Kumar, N et al. [18] proposed an Algorithm related to Security Encryption Algorithm for On- Line Analytical Processing (SEAOLAP) data cube. Kumar, N et al. [19] also proposed tangent and cotangent rule to generate the cypher and decipher text for E-communication. Awasthi. R et al. [20] have used two techniques clustering and outlier detection, which are used for fraudulent data and have proposed a method that takes less computation cost and perform better than the distance base method. Singh, P and Singh M. [21] have presented a proposed method for risk calculation by Bayesian Learning Approaches (BLA) based on risk score to find out whether transactions are fraudulent or genuine on two-parameter customer spending theBehaviour and geographical location if the risk score is 0.5 then transactions are considered fraudulent. Breed. D et al. [22] have presented a semi-supervised algorithm that is used to outperformance both an unsupervised and a supervised segmentation technique and compares with the Gini coefficient as a performance measure.

Dubey. R et al. [23] have applied machine learning unsupervised learning K-mean and different supervised learning method like K-NN, support vector machine, discriminate analysis and decision tree for data set authentication. Rajamohamed, R and Manokaran, J [24] have used a modified K-mean algorithm used for clustering and after that divides the clustered data into testing and training clusters after that classification sareperforms using a different algorithm and evaluates the accuracy, recall and misclassification error. Hassani, H et al. [25] have presented future development of both Data mining(DM) and the banking sector with a technique that provides the challenge are facing by the banking sector from big data. Dawood, E et al. [26] the author has studied the different algorithms like K-mean, improved K-mean, fuzzy C-mean and neural network by comparing the accuracy ratio are showed that the neural network is the best clustering technique. Yuan, C et al. [27] have presented the selection of K-values for the K-mean clustering algorithm and elbow method, canopy, silhouette Coefficient and Gap Statistics are used. Bai, C et al. [28] have used fuzzy C-mean clustering and fuzzy rough–set theory for finding the complex relation between farmer characteristics credit level, and competitive environment factor. Yang, M. S and Sinaga, K. P [29] have proposed Feature-Reduction Multi-View K-means (FRMVK), where weight feature and its computational complexity are also analysed.

Nazari, Aet al. [30] have proposed two-step K-mean and Kohonentechniques which are used for clustered and determination of the behaviour of resultant cluster and result shows that the SVM and Artificial Neural Network(ANN) are the most efficient techniques among eight different machine learning techniques used for clustering. Gupta, M. K and Chandra, P [31] have presented different data mining methods which are implemented in various real-life applications of data mining and discussed its challenge and issues. John Britto. M and Gobinath [32] have explored the data reduction technique through the K-mean clustering technique that enhances feature extraction. Bhatore, S et al. [33] have presented a review of 136 papers on credit risk evaluation of the paper, which was published between 1993 to 2019; and observed that ensemble and hybrid models with neural networks and SVM are implemented for credit scoring. Voican, O et al. [34] have given data analysis of big data like finance and banking institutions usingthe data mining method. Olaniyi, A. S et al. [35] have studied customer behaviour by using the clustering technique

and the clustered result was classified using Super Vector Machine (SVM) and found its accuracy 97% in the banking industry. Amoozad Mahdiraji. H et al. [36] analysed customer data using a K-mean clustering and modified recency, frequency and monetary (RFM) model and proposed a novel multi- attribute data mining approach and this approach has been implemented in large-scale problems in the financial service industry. Sivaguru,M and Punniyamoorthy. M [37] have proposed the weight optimization scheme for walandwuis to consider the presence of random choice of initial cluster centre for stability and show the improved performance of the RKM algorithm over clustering accuracy, convergence speed and evaluation metrics. Caruso. G et al. [38] have proposed a new methodology based on both quantitative and qualitative features for the estimation of credit risk. Gavira-Durón, N et al. [39] have proposed an analysis methodology based on the homogenous Markov chain and K-mean clustering method to determine the threshold and migration between clusters. Moslehi, F and Haeri, A [40] proposed K-mean cluster was created and examined on four data sets and compared its results then showed improved classification accuracy. Abbas, A et al. [41] have studied the use of unregulated learning clustering techniques on the employee of the production industry and implemented them in five manufacturing industries and research finding on the value of the Davis Bouldin Index (DBI).

Chandrasekaran,Get al. [42] described different challenges like content retrieval,content placement, stateful storage selection and proposed CASE-A Content-Aware Storage placement and retrieval Ecosystem. Mohammadian, V et al. [43] studied the gap between the fault-tolerance load balancing method in cloud computing. Hashem, I.A.T et al. [44] discussed the Hadoop technology, big data storage system and the relationship between big data and cloud computing and also focused on research challenges of scalability, data transportation, availability, data integrity, dataquality, data heterogeneity,privacy and regularity issues. Sosa-Sosa et al. [45] introduced methodology to improve the capacity utilisation and IO performance of cloud storage for Content Delivery and Sharing (CDS) service whereas the implementation of this method is done over private cloud for experimental evaluation. Chen,F et al. [46] focused on the solution of cloud storage security problems with the use of dynamic verifiable data access (DVDA). Liu, M et al. [47] proposed a cost-effective cloud-of cloud framework for two-tier storage with its cost to the different selection of tiers for data. Jin, Z et al. [48] proposed a Hoare- style proof system to clarify and analyse the Cloud Storage System (CSS) management program with the help of two languagesthe first one is modelling language and another one is an assertion language based on separation logic. Majed , Aet al. [49] proposed dynamic data relation strategy that is based on peer to peer cloud architecture. Xu, M et al. [50] proposed a new university student data management, which is baSsed on cloud storage technology. Chen, Y et al. [51] proposed a Cora, which is a policy-driven to support data correlation-based storage policy for cloud object storage. Wang, C et al. [52] analysed Chen et al.'S protocol regarding the security of cloud storage is insecure and provided an improved Usable Cloud Storage Auditing (UCSA). Yang, C et al. [53] reviewed the problem of dynamic insertion, integrity auditing and provable deletion over outsourced data and design a novel data structure called number- ranked-based Merkle has tree (NR-MHT), which is used in data integrity, auditing and dynamic data insertion.

3. PROPOSED METHOD

For making the database unique over the cloud by the use of K-mean technology, which is shown in given steps of algorithm below, and a flow diagram of proposed model is defined and represented in the following figure 2.

The K-means algorithm for partitioning, where each cluster centre are represented by the mean values of the objects in the cluster.

Input:

K: the number of cluster.

D: a data set containing n objects.

Output: A set of K clusters.

Method:

Step1: Select the number K to decide the number of clusters.

Step 2: Select random K points or centroids. (It can be other from the input dataset)

Step 3: Assign each data point to their closest centroid, which will form the predefined K

clusters

Step 4: Calculate the variance and place a new centroid of each cluster.

Step 5: Repeat the third steps, which means reassign, each data points to new closest centroid of each cluster.

Step 6: If any reassignment occur, then go to step 4 else go to finish.

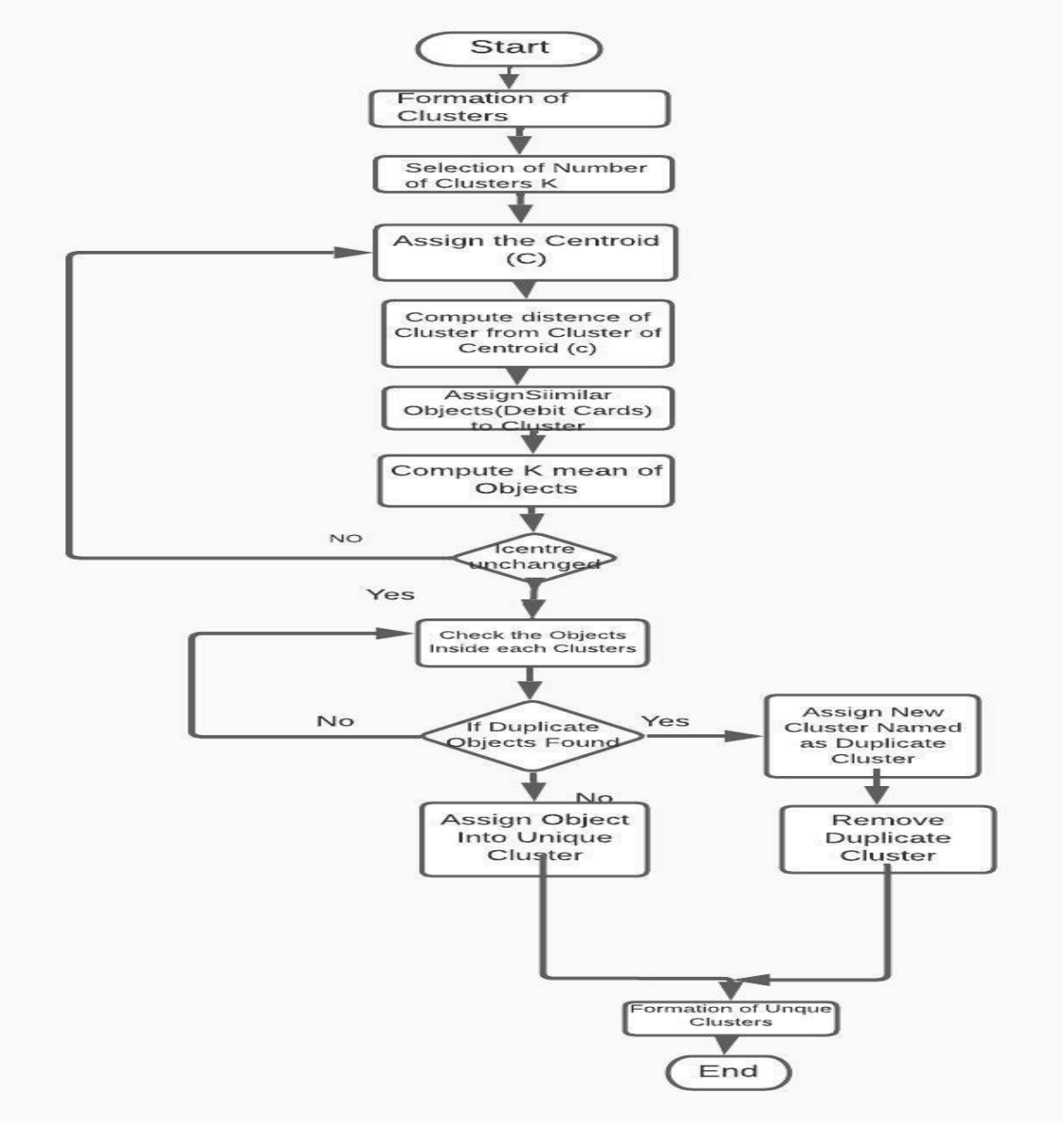


Figure 2. K-Means Clustering for Removal of Duplicate Clusters

In the present work, a set of debit cards is collected and the set of database may contain duplicacy for which K-mean technique is applied to identify the unique cluster, which helps to the users to search the desire information within minimum time. The following figure depicts the formation of unique cloud via K-means technology in which all cards are taken together to form a cluster and k clusters are identified are identified and a centroid is considered and on the basis of minimum distance of cluster from android ,extract the cluster and thereafter one can check the duplicate record which is placed into the duplicate record which is placed into the duplicate cluster and figure formation of unique cluster through through k-Means rest of the records are stored into the unique cluster and after deletion of duplicate cluster, users have only have only option to search into the unique cluster.

4. RESULTS AND DISCUSSION

In the above work, the data set of bankcards issued by bank and represented in table 1, and for clustering purpose, card number. As per data represented in the table, digits of card number is sum all together and considered as data point and shown below in table 2. The customer isrepresented by its Identification number along with as data point, and shown below in table2. The customer is represented by its Identification number along with tuples as Age, Card1, Card2 and Card3. The depicts of sum of all digits of cards are represented in table2.

Table 1. Representation of details of card Numbers

ID	Customer	Age	Card1	Card2	Card3
101	S1	52	4200778927351900	4492907363241340	4423916322437560
102	S2	32	4351245711358470	4835263951256210	4139636597873050
103	S3	44	4200101521325820	4682817921471080	4384717548838970
104	S4	38	4050453813353890	4503946857111580	4934193188723190
105	S5	35	4361664384555660	4426437886798500	4331593599718930
106	S6	28	4183026391245030	4009911666236670	4053467692969280
107	S7	25	4874657217235480	4690062165422420	4129564158434530
108	S8	43	4053057452828780	4086269975915990	4153727943447130
109	S9	36	4153727943447130	4367062511982550	4071956557833210
110	S10	27	4275688577132560	4800024388761150	4392132741528100
111	S11	45	4809729378151410	4247913936211590	4490293574881720
112	S12	42	4009162922114070	4750324891411770	4194291258997530
113	S13	30	4445466134688480	4828609892739930	4154868788586040
114	S14	47	4782495158244830	4853907545314310	4629317268614880
115	S15	53	4926851896119880	4414018673565270	4338473626563550
116	S16	55	4321402897333950	4366718211745060	4124422612745050
117	S17	50	4819825422548530	414824459383809	4079566211668550
118	S18	48	4476131559731100	4165021232865680	4811964665957910
119	S19	25	3733803757525570	4190874265258930	4510461921512960
120	S20	33	4449521734654070	4148815184118600	4148931417744020
121	S21	39	4843288947855130	3728929234194840	4472372596232140
122	S22	44	4906992218896500	4252335329317210	4471306376741890
123	S23	34	4758977475625990	4070622149611660	4085046159573380
124	S24	29	4912716221948540	4121866593879090	4471306376741890
125	S25	40	4605826661673470	4314679136824100	4038581676739020
126	S26	52	4200778927351900	4492907363241340	4423916322437560
127	S27	45	4809729378151410	4247913936211590	4490293574881720
128	S28	53	4926851896119880	4414018673565270	4338473626563550
129	S29	38	4050453813353890	4503946857111580	4934193188723190
130	S30	42	4009162922114070	4750324891411770	4194291258997530

Customer	Age	Card1	Card2	Card3
S1	52	73	64	66
S2	32	61	70	89
S3	44	42	74	91
S4	38	67	73	80
S5	35	80	86	84
S6	28	59	71	89
S7	25	65	59	67
S8	43	84	69	91
S9	36	65	68	72
S10	27	84	58	54
S11	45	77	73	81
S12	42	48	70	80
S13	30	82	96	86
S14	47	82	70	76
S15	53	94	68	79
S16	55	66	70	57
S17	50	70	72	72
S18	48	61	61	85
S19	25	76	82	75
S20	33	73	68	66
S21	39	79	83	65
S22	44	87	57	100
S23	54	101	60	76
S24	29	65	79	71
S25	40	75	62	78
S26	52	73	64	66
S27	45	77	73	81
S28	53	94	68	79
S29	38	67	73	80
S30	42	48	70	80

Table 2. Summing the Digits of cards

For clustering purpose, by taking the random value of k is consider as 3 and the first three – customers S1 S2 and S3 are taken as seeds of the respective clusters and it is represented in table 3.

Table 3. Representation of Seed Values

Customer	Age	Card1	Card2	Card3
S1	52	73	64	66
S2	32	61	70	89
S3	44	42	74	91

Let us perform the iterations and in this case the distance from centroid is computed and all the results are given in the table 4.

C1	42.07	79.07	65.84	72.84	From C1	From C2	From C3	Allocation to the nearest cluster
C2	34.66	70.77	78.55	80.77				
C3	44.66	50.33	68.33	85.33				
S1	52	73	64	66	24.68	48.89	53.67	C1
S2	32	61	70	89	48.46	29.21	28.67	C2
S3	44	42	74	91	65.32	60.82	35.82	C3
S4	38	67	73	80	30.46	13.43	23.33	C2
S5	35	80	86	84	39.32	20.25	48.11	C2
S6	28	59	71	89	55.46	34.21	31.67	C3
S7	25	65	59	67	43.82	48.75	61.99	C1
S8	43	84	69	91	27.18	41.35	41.67	C1
S9	36	65	68	72	24.14	26.43	36.99	C1
S10	27	84	58	54	46.68	68.21	92.99	C1
S11	45	77	73	81	20.32	22.35	15.57	C3
S12	42	48	70	80	42.46	39.43	11.99	C3
S13	30	82	96	86	58.72	38.57	74.67	C2
S14	47	82	70	76	15.18	36.89	45.01	C1
S15	53	94	68	79	34.18	53.89	58.67	C1
S16	55	66	70	57	46	57.43	56.01	C1
S17	50	70	72	72	24	31.43	42.01	C1
S18	48	61	61	85	41	44.89	21.67	C3
S19	25	76	82	75	38.46	24.11	69.33	C2
S20	33	73	68	66	24.14	29.21	53.99	C1
S21	39	79	83	65	28.14	32.79	69.33	C1
S22	44	87	57	100	45.86	66.35	63.33	C1
S23	54	101	60	76	42.86	72.89	77.67	C1
S24	29	65	79	71	42.14	21.65	55.33	C2
S25	40	75	62	78	15.14	28.89	42.99	C1
S26	52	73	64	66	24.68	48.89	53.67	C1
S27	45	77	73	81	20.32	22.35	15.57	C3
S28	53	94	68	79	34.18	53.89	58.67	C1
S29	38	67	73	80	30.46	13.43	23.33	C2
S30	42	48	70	80	42.46	39.43	11.99	C3

Table 4. Computation of Distance from Centroid

C1	52	73	64	66	From C1	From C2	FromC3	Allocation to the nearest Cluster
C2	32	61	70	89				
C3	44	42	74	91				
S1	52	73	64	66	0	62	74	C1
S2	32	61	70	89	61	0	37	C2
S3	44	42	74	91	74	37	0	C3
S4	38	67	73	80	50	24	43	C2
S5	35	80	86	84	59	43	66	C2
S6	28	59	71	89	68	07	38	C2
S7	25	65	59	67	41	44	81	C1
S8	43	84	69	91	50	37	48	C2
S9	36	65	68	72	34	27	56	C2
S10	27	84	58	54	54	75	112	C1
S11	45	77	73	81	35	40	47	C1
S12	42	48	70	80	45	32	23	C3
S13	30	82	96	86	83	52	81	C2
S14	47	82	70	76	30	49	62	C1
S15	53	94	68	79	41	68	81	C1
S16	55	66	70	57	25	60	73	C1
S17	50	70	72	72	19	46	55	C1
S18	48	61	61	85	38	29	42	C2
S19	25	76	82	75	57	48	77	C2
S20	33	73	68	66	23	38	73	C1
S21	39	79	83	65	39	62	77	C1
S22	44	87	57	100	63	62	71	C2
S23	54	101	60	76	44	85	98	C1
S24	29	65	79	71	51	34	63	C2
S25	40	75	62	78	28	41	62	C1
S26	52	73	64	66	0	62	74	C1
S27	45	77	73	81	35	40	47	C1
S28	53	94	68	79	41	68	81	C1
S29	38	67	73	80	50	24	43	C2
S30	42	48	70	80	45	32	23	C3

Table 5. Computation of New Mean Values

Proceed on the minimum distance, which is selected from above table, new mean is computed and represented in the table5.

In the similar manner, further iteration ae performed and a resultant cluster table is computed and represented in the table6.

Table 6. Representation of Resultant Cluster

	Age	Card1	Card2	Card3
C1	43.33	78.25	67.25	69.75
C2	35.27	71.54	73.81	83.81
C3	43	45	52	85.5
Seed1	52	73	64	66
Seed2	32	61	70	89
Seed3	44	42	74	91

From the above table, it is observed that by the use of K-means technique, theclusterC1 has the 16 customers S1, S7, S8, S9, S10, S14, S15, S16, S17, S20, S21, S22, S23 S25, S26, S28 whereas cluster C2 has 7 (seven) customers S2, S4, S5, S13, S19, S24, S29 and in cluster C3 has 7 (seven) customers S3, S6, S11, S12, S18, S27, S30. The above clusters are unique and duplicate clusters have been removed during computation, which is implementation through Python language. Let us consider the five attributes like ID,Age, Card1, Card2 and Card3 which are given in table 2. Which is extracted into two parts containing unique and duplicate records and is represented in the following figure 3.

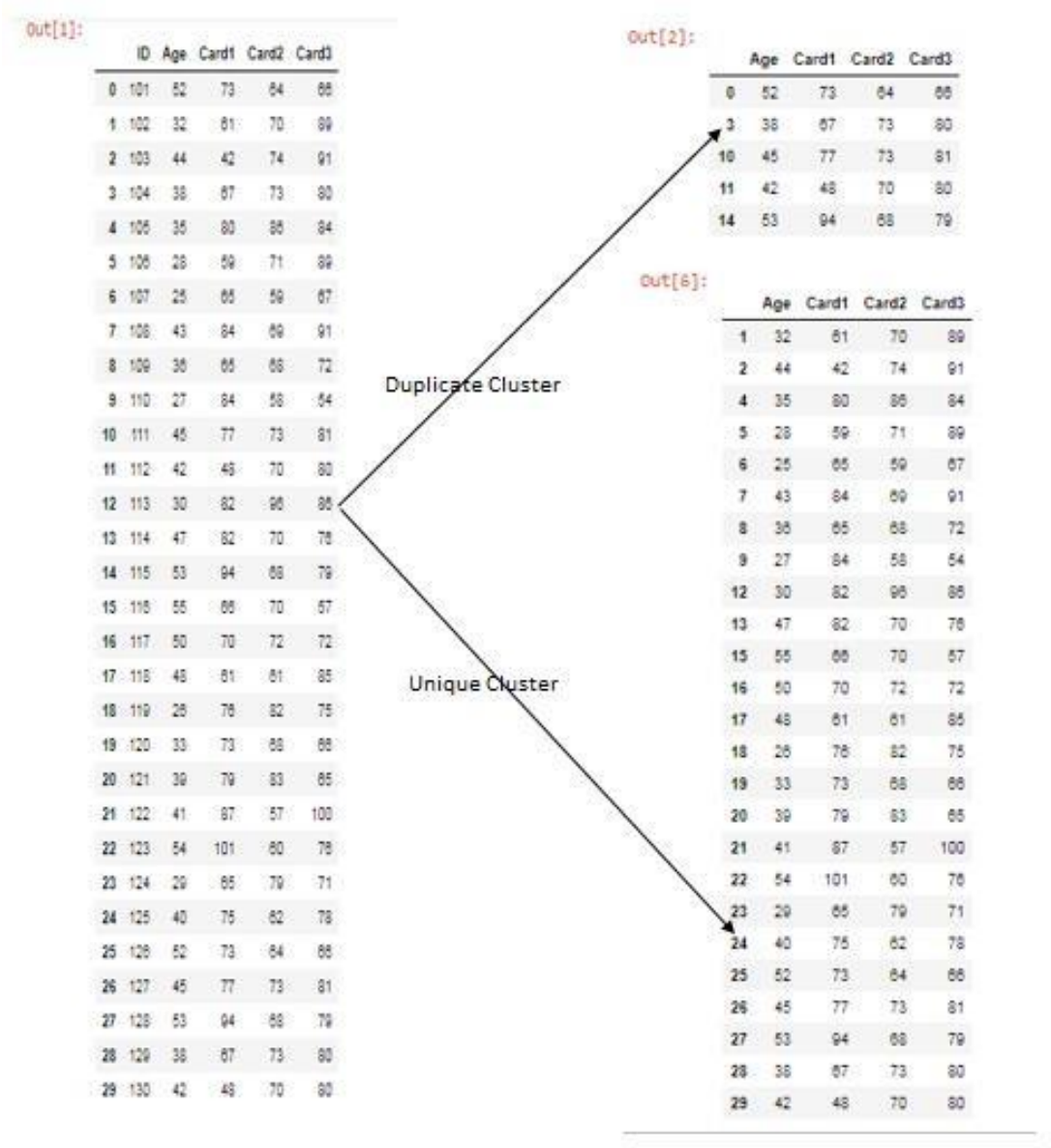


Figure 3. Breakdown of Sample Data Set into Unique and Duplicate Records

5. CONCLUSIONS

From the above work, it is concluded that the k-means is a sufficient technique that can be used for removal of duplicate records forming the duplicate cluster and it is useful over the cloud, which lies large number of data storage in the form of text, image, audio and video files. The above work can be extended for computation of space and time complexities for storing the cloud storage and accessing the datasets with in minimum time.

REFFERENCE

[1] C.J. Matheus,P.K.Chan, and G. Piatetsky-Shapiro(1993), "Systems for Knowledge Discovery in Databases" *IEEE Transactions on Data Engineering,* Vol. 5, No. 6, pp. 903- 913.

[2] Kevin Zhu, Kenneth L. Kraemer and Jason Dedrick(2004),"Information Technology Payoff in E-Business Environments: An International Perspective on Value Creation of E-Business in the Financial Services Industry" International Journal of Management Information Systems, Vol.21, No1, pp 17-54.

[3] Yu-Ju-Lin,Chin-Sheng, Huang, and Che-ChernLin (2008), "Determination of insurance policy using neural networks

and simplified models with factor analysis technique", International Journal WSEAS Transactions on Information Science and Applications, Vol. 5, No 10, pp 1405-1415.

[4] Shu-Hsien Liao,Ya-Ning Chen and Yu-YiaTseng (2008), "Mining Demand Chain Knowledge of Life Insurance Market for New Product Development" International Journal Expert Systems with Applications, Vol.36, No5, pp 9422-9437.

[5] Hurtado, C., Mendelzon, A. and Vaisman, A (2011), "Maintaining Data Cubes under Dimension Updates, Proc IEEE/ICDE '99 AnkitGupta, Arpit Gupta Amit Mishra "Research Paper On Cluster Techniques Of Data Variations", International Journal of Advance Technology & Engineering Research (IJATER), ISSN NO: Vol.1, No.1, (2011), pp 2250-3536.

[6] JayantiRanjan, Raghuvir Singh and Vishal Bhatnaga(2011), "Analytical Customer Relationship Management in Insurance Industry using Data Mining: a case Study of India Insurance company" International Journal of Networking and Virtual Organizations, Vol. 9, No. 4, pp. 331-366.

[7] Shu-hsien Liao, Yin-ju Chen and Yi-tsunLin(2011), "Mining Customer Knowledge to Implement Online Shopping and Home Delivery for Hypermarkets" International Journal Expert Systems with Applications, Vol. 38, No. 4, pp. 3982-3991.

[8] FaribaSadri (2012), "AmbientIintelligence: A survey" International Journal ACM Computing Survey(CSUR), Vol. 43, No. 4, Article No. 36.

[9] Ritu Sharma, M. AfsharAlam and Anita Rani (2012), "K-Means Clustering in Spatial Data Mining using Weka Interface", International Conference on Advances in Communication and Computing Technologies (ICACACT) Proceedings published by International Journal of Computer Applications® (IJCA).

[10] Shanshan Li, Bing Zhang andAnLi(2012), "XiupingJia; Hyper spectral Imagery Clustering withNeighborhood Constraints", Geoscience and Remote Sensing Letters, IEEE Transactions on, Vol. 10, No. 3, pp. 588 – 592.

[11] Mauricio C. Moraes, Carlos A. Heuser and Viviane P. Moreira (2012), "Pre-Query Discovery of Domain-specific Query Forms: A Survey" IEEE Trasactions on Data Engineering, ISSN: pp. 1041-4347.

[12] Mousazadeh, S and Cohen, "Audio, Speech, Language Processing, and I (2013), "Voice Activity Detection in Presence of Transient Noise Using Spectral Clustering", IEEE Transactions, Vol. 21, No.6, pp. 1261 – 1271.

[13] Djidjev, Hristo N and Onus, Melih (2013), "Scalable and Accurate Graph Clustering and Community Structure Detection", Parallel and Distributed Systems, IEEE Transactions, Vol. 24, No.5, pp.1022 - 1029.

[14] Papalexakis, E.E., Sidiropoulos, N.D and Bro, R (2013), "From K-Means to Higher- Way Co-Clustering: Multilinear Decomposition with Sparse Latent Factors", Signal Processing, IEEE Transactions, Vol. 61 No. 2, pp.493 – 506.

[15] Ali, M., Ilie, I.-S.; Milanovic, J.V.; Chicco, G. (2013), "Wind Farm Model Aggregation Using Probabilistic Clustering Power Systems", IEEE Transactions, Vol. 28, No.1, pp.309 – 316.

[16] Joseph R. Kasprzyk, ShanthiNataraj and Patrick M. Reed (2013), "Many Objective Robust Decision Making for Complex Environmental Systems undergoing Change" International Journal Environmental Modeling& Software, Vol. 42, pp. 55-71.

[17] Kumar, N., Verma, V., and Saxena, V (2013), "Cluster analysis in data mining using k- means method". International Journal of Computer Applications, Vol. 76, No. 12, pp.11-14.

[18] Kumar, N., Verma, V. andSaxena, V (2013), "A security algorithm for online analytical processing data cube", International Journal of Computer Applications, Vol. 79, No.14.

[19] Kumar, N., Verma, V. andSaxena, V.

(2013), "Implementation of Tangent and Cotangent Rules to Provide Security in E-communication". International Journal of Computer Applications, 975, 8887.

[20] Awasthi, R., Tiwari, A. K., and Pathak, S (2013), "Empirical Evaluation On K Means Clustering with Effect of Distance Functions for Bank Dataset". *International Journal of Innovative Technology and Research*, Vol. 1, No.3, pp. 233–235.

[21] Singh, P., and Singh, M (2015), "Fraud Detection by Monitoring Customer Behavior and Activities". *International Journal of Computer Applications*, Vol.*111*, No.11. https://doi.org/10.5120/19584-1340

[22] Breed, D., Verster, T., and Terblanche, S. A (2017), "Semi-Supervised Segmentation Algorithm as Applied to K-means Using Information Value". *ORiON*, Vol, *33*, No.2. https://doi.org/10.5784/33-2-568

[23] Dubey, R. K., and Nizar Banu, P. K. (2018), "Analysis of Supervised and Unsupervised Technique for Authentication dataset". *International Journal of Engineering and Technology(UAE)*, vol.*7, No.*4. https://doi.org/10.14419/ijet.v7i4.9564

[24] Rajamohamed, R., and Manokaran, J. (2018), "Improved Credit Card Churn Prediction Based on Rough Clustering and Supervised Learning Techniques". *Cluster Computing*, Vol.*21, No.*1. https://doi.org/10.1007/s10586-017-0933-1

[25] Hassani, H., Huang, X., and Silva, E (2018), "Digitalisation and Big Data Mining in Banking". In Big Data and Cognitive Computing, Vol. 2, No.3, pp.1–13. MDPI AG. https://doi.org/10.3390/bdcc2030018

[26] Dawood, E. A. E., Elfakhrany, E., and Maghraby, F. A (2019), "Improve profiling bank customer's behavior using K-means". *J*, Vol.*2, No.*2,

[27] Yuan, C., and Yang, H (2019), "Research on K-Value Selection Method of K-MeansClustering Algorith

[28] Bai, C., Shi, B., Liu, F., and Sarkis, J (2019), "Banking credit worthiness: Evaluating the complex relationships". *Omega (United Kingdom)*, Vol. *83*, pp.26–38. https://doi.org/10.1016/j.omega.2018.02.001

[29] Yang, M. S., and Sinaga, K. P. (2019)," A Feature-Reduction Multi-View K-means Clustering Algorithm". *IEEE*

[30] Nazari, A., Mehregan, M., and Tehrani, R (2020),"Evaluating the Effectiveness of Data Mining Techniques in Credit Scoring of Bank Customers Using Mathematical Models: A Case Study of Individual Borrowers of Refahkargaran Bank in Zanjan Province, Iran". *International Journal of Nonlinear Analysis and Applications*, Vol.11(Special Issue). https://doi.org/10.22075/ijnaa.2020.4604

[31] Gupta, M. K., and Chandra, P. (2020), "A comprehensive survey of data mining". *International Journal of Information Technology (Singapore)*, Vol. *12, No.*4, pp.1243–1257. https://doi.org/10.1007/s41870-020-00427-7

[32] John Britto, M., and Gobinath, R (2020), "REHC: Reduction through exclusive homogeneous clusters for imbalance dataset". *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2020.11.093

[33] Bhatore, S., Mohan, L., and Reddy, Y. R. (2020), "Machine Learning Techniques for Credit Risk Evaluation: A Systematic Literature Review". *Journal of Banking and Financial Technology*, Vol.4, No.1, pp. 111–138. https://doi.org/10.1007/s42786-020-00020-3

[34] Voican, O (2020), "Using Data Mining Methods to Solve Classification Problems in Financial-Banking Institutions". *Economic Computation and Economic Cybernetics Studies*

[35] Olaniyi, A. S., Olaolu, A. M., Jimada-Ojuolape, B., and Kayode, S. Y (2020), "Customer Churn Prediction in Banking using Machine Learning". *IEEE Access*, *and*

Industry Using K-Means and Support Vector Machine Algorithms". *International Journal of Multidisciplinary Sciences and Advanced Technology*, Vol.1, No.*1*, pp. 48–54.

[36] AmoozadMahdiraji, H., Tavana, M., Mahdiani, P., and AbbasiKamardi, A. A (2021)," A Multi-Attribute Data Mining Model for Rule Extraction and Service Operations Benchmarking". *Benchmarking*. https://doi.org/10.1108/BIJ-03-2021-0127

[37] Sivaguru, M., and Punniyamoorthy, M (2021), "Performance-Enhanced Rough K- means Clustering Algorithm". *Soft Computing*, Vol. *25, No.*2, pp. 1595–1616. https://doi.org/10.1007/s00500-020-05247-2

[38] Caruso, G., Gattone, S. A., Fortuna, F., and Di Battista, T (2021)," Cluster Analysis for mixed data: An application to credit risk evaluation". *Socio-EconomicPlanning Sciences*, Vol.*73*,100850. https://doi.org/10.1016/j.seps.2020.100850

[39] Gavira-Durón, N., Gutierrez-Vargas, O., and Cruz-Aké, S. (2021), "Markov Chain K- means Cluster Models and Their Use for Companies' Credit Quality and Default Probability Estimation". *Mathematics*, Vol.*9, No.*8,

[40] Moslehi, F., and Haeri, A. S (2021)," A Novel Feature Selection Approach Based on Clustering Algorithm". *Journal of Statistical Computation and Simulation*, Vol. *91, No. 3*, pp. 581–604.https://doi.org/10.1080/00949655.2020.1822358

[41] Abbas, A., Prayitno, P., Butarbutar, F., Nurkim, N., Prumanto, D., Dewadi, F. M., Hidayati, N., and Windarto, A. P (2021)," Implementation of clustering unsupervised learning using K-Means mapping techniques". *IOP Conference Series: Materials Science and Engineering*, Vol. *1088, No.1*, 012004. https://doi.org/10.1088/1757-899x/1088/1/012004

[42] Chandrasekaran, G., Ramasamy, B., Dhondi, P., Thorat, P. B.andChalla, R (2022), "CASE: A Context-Aware Storage Placement and Retrieval Ecosystem". IEEE Access, Vol.10. https://doi.org/10.1109/ACCESS.2021.3139339

[43] Mohammadian, V., Navimipour, N. J., Hosseinzadeh, M., and Darwesh, A (2022), "Fault-Tolerant Load Balancing in Cloud Computing: A Systematic Literature Review". IEEE Access, Vol.10. https://doi.org/10.1109/ACCESS.2021.3139730

[44] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A.andUllah Khan, S (2022), "The Rise of "Big Data" on Cloud Computing: Review and Open Research Issues". In Information Systems, Vol.47. https://doi.org/10.1016/j.is.2014.07.006

[45] Sosa-Sosa, V. J., Barron, A., Gonzalez-Compean, J. L., Carretero, Jand Lopez-Arevalo, I. (2022)," Improving Performance and Capacity Utilization in Cloud Storage for Content Delivery and Sharing Services". IEEE Transactions on Cloud Computing, Vol.10, No.1. https://doi.org/10.1109/TCC.2020.2968444

[46] Chen, F., Li, Z., Jiang, C. and Li, J. (2022)," Verifiable Cloud Data Access Design, Analysis, and Implementation". IEEE Systems Journal, Vol.16, No.1. https://doi.org/10.1109/JSYST.2020.3034105

[47] Liu, M., Pan, L. and Liu, S. (2022)," Effeclouds: A Cost-Effective Cloud-of-Clouds Framework for Two-Tier Storage". Future Generation Computer Systems, 129. https://doi.org/10.1016/j.future.2021.11.012

[48] Jin, Z., Zhang, B., Zhang, L., Cao, Y., and Wang, H. (2022), "An Adaptation-Complete Proof System for Local Reasoning about Cloud Storage Systems". Theoretical Computer Science, 903. https://doi.org/10.1016/j.tcs.2021.12.018

[49] Majed, A., Raji, F and Miri, A. (2022), "Replication management in peer-to-peer cloud storage systems". Cluster Computing, vol. 25, no. 1.https://doi.org/10.1007/s10586-021-03395-0

[50] Xu, M. and Liu, Y (2022), "Achievement Management System for University Students Based on Cloud Storage Technology". International Journal of Information and Communication Technology, Vol.20, No. 1. https://doi.org/10.1504/IJICT.2022.119312

[51] Chen, Y., Tong, W., Feng, D. and Wang, Z. (2022)," Cora: Data Correlations-Based Storage Policies for Cloud Object Storage". Future Generation Computer Systems, Vol.129. https://doi.org/10.1016/j.future.2021.11.023

[52] Wang, H., Feng, L., Ji, Y., Shao, B. and Xue, R (2022), "Toward Usable Cloud Storage Auditing, Revisited". IEEE Systems Journal, Vol.16, No. 1. https://doi.org/10.1109/JSYST.2021.3055021.

[53] Yang, C., Liu, Y., Zhao, F. and Zhang, S. (2022), "Provable data deletion from efficient data integrity auditing and insertion in cloud storage". Computer Standards and Interfaces, Vol.82. https://doi.org/10.1016/j.csi.2022.103629.