Adaptive Hierarchical Clustering using Hashing

[1]Kirti Kshirsagar, [2]Chetan Patil, [3]Anjali Mahajan
[1]BE, [2]BE, [3]BE
[1]Information Technology,
[1]MET Bhujbal Knowledge City, Nashik India.

Abstract: Here, the system shows an design of Adaptive Hierarchical Clustering in which, any type of data, let it be structured or unstructured data. The streaming data will be stored in database using operations like hash key generation and clustering of data using map-reduce. To show this in more understanding manner, system will be using theme of hospital, in which the data of patients and their diseases will be uploaded and using SHA algorithm hash key of particular file will be generated and like map-reduce framework that file will be divided into clusters and stored in database. Other user can access it easily as per requirements.

IndexTerms - Map-reduce, Hierarchical Clustering, Adaptive Clustering.

I. INTRODUCTION

As research show in existing clustering method we get multiple clusters that means we get n-number of clusters for similar data so to avoid this we are creating system for given data. Although functional distributed computation framework have seen increased adoption in both research and software development communities, standard libraries such as Apache SparkMLib currently offer only simple, parametric clustering methods.Sophisticated clustering approaches on such platform has become increasingly pressing due to modern instrumentation and new technologies, such as the Internet Of Things (IoT). New clustering implementations that leverage this architecture are desirable.

Our contributions square measure a theoretical discussion on the protection of a medical info. The implementation of a prototype to simulate a secure medical database and the result of several experiments that we conducted. In this thesis, we show that the performance impact of providing confidentially and integrity within a medical database is considerable. Even though our model is comparatively slow, in practice the impact is probably less. If a doctor has to wait the non-secure medical database, the security benefits will outweigh the performance impact. This ends up in the conclusion that firmly planning a medical information is feasible while not swing trust within the info itself. The stored data will be secured by using SHA algorithm it will create an hash-key and stored in database after its clustering is done so authorized user can access data as per need.

II. RELATED WORK

With the emergence of big data and cloud computing, data stream arrive rapidly, large-scale and continuously, real time data stream clustering analysis has become a hot topic in the study on the current data stream minminin. Some existing data stream clustering algorithm cannot effectively deal with the high dimensional data stream and are incompetent to find clusters of arbitrary shape in real time, as well as noise points could not be removed timely.As in existing clustering method we get multiple clusters that mean we get n-number of clusters for similar data.System use Map-reduce framework and it can be used on any platform. Clustering is done using Map-reduce method Recently, data collected from business have continuously growing in every enterprise. The Big Data, Cloud Computing, Data Mining has become hot topics at the present day. How to acquire important information quickly from these data is a critical issue. In this paper, we modified the traditional A priori algorithm by improving the execution efficiency, since Aprori algorithm has confronted with a drawback that the computation time increases dramatically when data size increases. Since the one-phase algorithm only used one MapReduce operation, it will generate excessive candidates and result in insufficient memory[1]. Clustering is an important phase in data mining. A number of different clustering methods are used to perform cluster analysis: Partitioning Clustering, hierarchical clustering, grid-based clustering, model-based, graph based clustering and density based clustering and so on. Hierarchical method helps us to cluster the data objects in the form of a tree known as hierarchy. And each node in hierarchy is known as the cluster. Hierarchical clustering can be performed in two ways: agglomerative clustering and divisive clustering. Agglomerative clustering is always more preferable. For a good cluster analysis, the quality of the clusters should be high. In this paper, we will measure the quality of clusters with the help of three parameters: Cohesion measurement, Silhouette index and Elapsed time[2]. Cloud computing is more popular for storing information by using storing statistics. By storing data into the cloud save the economical funds and gives the great flexibility. Data safety is the most challenging for the researchers. Confidential information must be encoded before outsourcing the data. Hierarchical clustering method presented in this paper used to the privacy preserving powerful search over encoded cloud data. This search over outsourced information, security is considered as a user revocation approach. Another more important component of this framework is the data duplication and it checking using SHA1 hashing strategy. Strategy will not allow saving the replica data at cloud server. EM clustering uses for clustering process and compare the EM clustering algorithm with K-means clustering

algorithm. Experimental results state that the framework has various advantages like time utilization, efficient, easy and statistics duplication checking[3].Web services (WS) is called composite or compound when its execution involves interactions with other WS to utilize their features. The service providers published the web services through the internet as independent software components that are fulfilling the requirements of customer request. Clustering is more necessary for efficient web service discovery and web service composition processes. Clustering process groups the similar type of web services. In this paper, efficient clustering methods such as k-means clustering, Hierarchical agglomerative clustering and Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) clustering are introduced for web service clustering. The k-means clustering is a kind of partitional clustering where the web pages are divided into subsets with no hierarchy defined over them and the hierarchical agglomerative clustering is a type of hierarchical clustering where the web pages are arranged in tree structure in which leaves represents the data points and nodes denotes the clusters. BIRCH is an integrated hierarchical clustering algorithm uses the clustering features and cluster feature tree for general cluster description. Based on these clustering methods, web pages are clustered which are used for web service discovery and web service composition. The experiments are conducted on number of web services and the efficiency is evaluated in terms of accuracy, precision, recall and run time [4]. Generally, the medical datasets are heterogeneous and large dimensional that contains a million of patient records. Extracting information from such datasets is a tedious process, which can be made easier by some of the clustering algorithms available in data mining. In this paper, three clustering algorithms such as Medical Storage Platform for data Mining (MSPM), Homogeneity Similarity based Hierarchical (HSH) clustering and K-Harmonic Means-Overlapped K-Means (KHM-OKM) clustering is described and their performance is evaluated. The HSH clustering is an enhancement of hierarchical algorithm that considers the homogeneity and relative population of the clusters to measure the Clustering Performance Index (CPI). The MSPM framework is a modification of Apriori algorithm implemented using MapReduce function. This framework enhances the performance of clustering by the parallel processing of Map and Reduce functions. The KHM-OKM clustering is a hybrid algorithm that combines the K-Harmonic Means and Overlapped K-Means clustering. The results of these algorithms are experimentally evaluated regardingCPI, confidence and FBCubed measure[5]. We present an accelerated algorithm for hierarchical density based clustering. Our new algorithm improves upon HDBSCAN*, which itself provided a signi?cant qualitative improvement over the popular DBSCAN algorithm. The accelerated HDBSCAN* algorithm provides comparable performance to DBSCAN, while supporting variable density clusters, and eliminating the need for the dif?cult to tune distance scale parameter This makes accelerated HDBSCAN* the default choice for density based clustering [6].

III. AIM AND OBJECTIVE'S

The aim of this project is to demonstrate hashing function on adaptive hierarchical clustering with the help of map-reduce which is used for clustering with the help of map-reduce framework.

3.1 Objective's

- To design clustering system for data cleaning.
- To improve the efficiency of clustering using batch processing.
- To design efficiency hierarchical cluster of data.

IV. MOTIVATION

The necessity to introduce for effective search and clustering Although functional distributed computation frameworks have seen increased adoption in both research and software development communities, standard libraries such as Apache SparkMLib currently offer only simple, parametric clustering methods. The need for more sophisticated clustering approaches on such platforms has become increasingly pressing due to modern instrumentation and new technologies, such as the Internet of Things (IoT). So new clustering implementations that leverage this architecture are desirable. In addition, data often arrive in the form of unstructured text logs, rather than numerate or enumerated structured data.

V. System Architecture

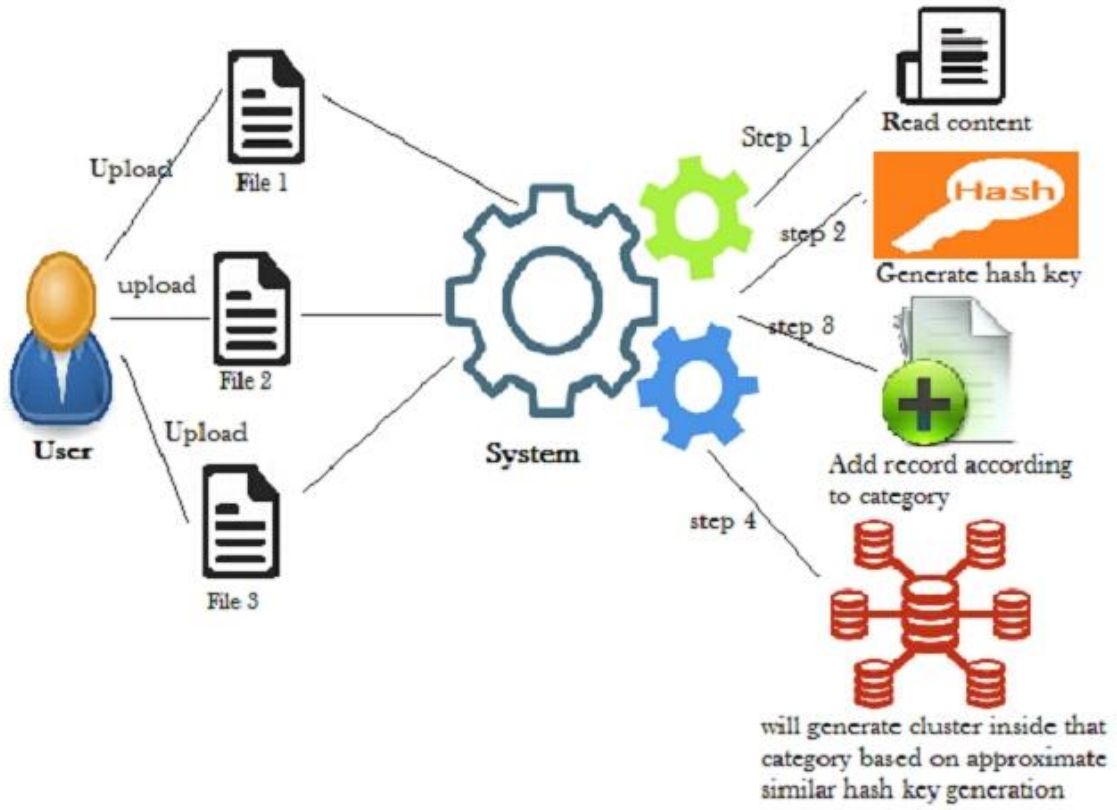

Figure 1. Shows Architecture of System

In system architecture it is shown that user means Doctor upload data on system and system will generate hash-key of data after completion of hash-key Clustering Algorithm will be performed on the data and will be stored in database.

Following are the Algorithms used to carry out requirements of System

(a) Secure Hash Algorithm In this system SHA-1 algorithm is used it is cryptography hash function which takes an input and produces a 160-bit(20-byte) hash value known as a message digest typically rendered as hexadecimal number,40 digit long. It was designed hierarchical clustering

(b) Hierarchical clustering,also known as hierarchical cluster analysis,is an algorithm that groups similar objects into groups called clusters. The terminus could be a set of clusters,where each cluster is distinct from each other cluster,and the objects within each cluster are broadly similar to each other. Hierarchical cluster starts by treating every observation as a separate cluster.

Then, it repeatedly executes the following two steps,

- Identify the two clusters that are closest together.
- Merge the most similar clusters

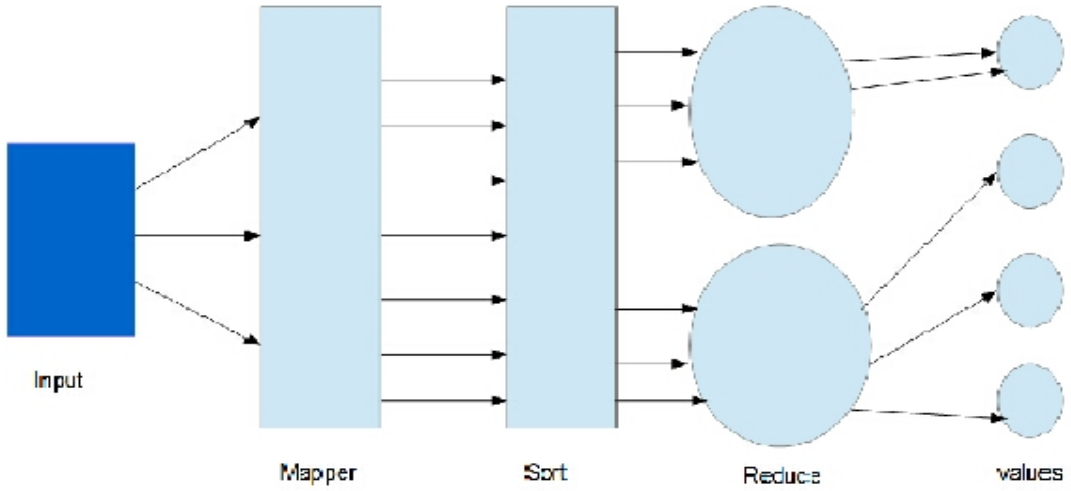

Figure 2:- Map-reduce Frame work.

Map reduce may be a framework for process parallelizable issues across massive datasets employing a sizable amount of computers (nodes), together said as a cluster (if all nodes square measure on constant native network and use similar hardware) or a grid (if the nodes unit of measurement shared across geographically and administratively distributed systems, and use tons of heterogeneous hardware). Processing can occur on data stored either in a file system(unstructured)or in a data system(structured). Map scale back will make the most of the neighbourhood of information, processing it near the place it is stored in order to minimize communication overhead.

A Map scale back framework (or system) is sometimes composed of 3 operations (or steps):

• Map: A lot of workers apply the function to map the local data and write the output temporarily. There is also a master node which ensures the redundant data produces only one copy

• Shuffle: worker nodes redistribute data based on the output keys (produced by the map function), such that all data belonging to one key is located on the same worker node.

• Reduce: The worker nodes also process individual groups of output data, per key, and in parallel.

Following are the main functions used in this System

•SHA

•Clustering

Hash functions area unit very helpful and seem in the majority data security applications. A hash operate may be function that converts a numerical input worth into another compressed numerical worth. The input to the hash operate is of absolute length however output is usually of fastened length. Values came by a hash perform area unit known as message digest or just hash values. SHA-1 is that the most generally used of the present SHA hash functions. It is used in many wide used applications and protocols as well as Secure Socket Layer (SSL) security. Clustering is dividing information points into undiversified categories or clusters, when collection of objects is given, we put objects into group based on similarity. If k is given, the K-means algorithmic program will be dead within the following steps:

• Partition of objects into k non-empty subsets.

• Identifying the cluster centroids (mean point) of the present partition.

• Assigning each point to a specific cluster.

•Compute the spaces from every purpose and allot points to the cluster wherever the distance from the centre of mass is minimum.

• After re-allotting the points, notice the centre of mass of the new cluster fashioned.

4.1 **Mathematical Formula** for K-mean

D={x1,x2,....,xi,....,xm}data set of m records

.xi= (xi1,xi2,.....,xin) each record is an n-dimensional vector.

$$C_j = Cluster(x_i) = \arg_j min \| x_i - u_j \|^2$$

$$Distortion = \sum_{n=1}^{m} (x_i - c_i)^2 = \sum_{n=1}^{k} \sum_{ieOwnedby(\mu_j)} (x_i - \mu_i)^2 /$$

- Finding Cluster Centres that Minimize Distortion:
  Solution can be found by setting the partial derivative of Distortion w.r.t.each cluster center to zero.

$$\frac{\text{Distortion}}{\mu_i} = \frac{}{\mu_j} \sum_{ieOwnedby(\mu_j)} (x_i - \mu_i) = -2 = \frac{\text{Distortion}}{\mu_i} = \frac{}{\mu_j} \sum_{ieOwnedby(\mu_j)} =$$

$$(Forminimum)$$

$$(\mu_j) = \frac{1}{\|Ownedby(\mu_j)\|} \sum_{ieOwnedby(\mu_j)} x_i$$

k-means clustering with example For any k clusters, the value of k should be such that even if we increase the value of k from after several levels of clustering the distortion remains constant. The achieved point is called the This is the ideal value of k, for the clusters created. K-means clustering is used for clustering system data.

VI. CONCLUSION

The clustering method being used by practitioners on such systems do not respond rapidly to new data, or do not adjust the number of clusters appropriately as more data, or do not adjust the number of clusters appropriately as more data are processed. This problem is particularly acute for unstructured data like text and other non-enumerated types that are common in log and message stream and not analyzed at scale for precisely this reason. To address such issue this system use a method for hierarchical clustering using adaptive hash(AdaHash) values that can be recalculated during a periodic batch process and used for subsequent streaming processing at the speed of data arrival, assuming sufficient distributed compute resources. this system is as fast as other optimal hashing function with using suitable farm work

This work presents Adaptive Hierarchical clustering on data, for clustering large amount of data without any glitch. Also without over head of similar cluster. Clustering is done on hash key generated data, This will make system efficient and not make similar type of clustering.

VII. ACKNOWLEDGEMENT

It gives us great pleasure to acknowledgement our deep sense of gratitude towards our respected Prof.Priti Lahane.For her valuable guidance, profound advice, persistent encouragement and help during the completion of this work.Her time to time suggestions boosted us to complete this task successfully. She has helped us in all possible way right from gathering information to report presentation. We express our thanks to our project coordinator

Dr.Kalpana Metre for her kind cooperation.We extend our sincere thanks to our Head Of Department DR.S.V.Gumaste for providing all kinds of cooperation during course.Last but not least we are very grateful for the staff members and people, who helped us directly or indirectly for this project. Thank you.

REFERENCES

1. International Journal of Data Science and Analytics AdaHashh: hashing-based scalable, adaptive hierarchical clustering of streaming data on Map-reduce framework" year 2018 by Dean Teffer, Ravi Shastri, Joydeep Ghosh.

2. 1st International Conference on Next Generation Computing Technologies (NGCT)"Cluster quality based performance evaluation of hierarchical clustering method" year 2015 by Nisha, Punit Jai Kaur.

3. Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) "Clustering Based Efficient Privacy Preserving Multi Keyword Search Over Encrypted Data" year 2018 by Neha Mahajan ; Vaishali Barkade.

4. IEEEIEEE International Conference on Computational Intelligence and Computing Research (ICCIC)"Efficient Clustering Techniques for Web Services Clustering"year 2017 by T Parimalam ; K Meenakshi Sundaram.

5. Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)"Performance analysis of clustering algorithms in medical datasets" year 2017 by P. Premalatha ; S. Subasree.

6. IEEE International Conference on Data Mining Workshops "Accelerated Hierarchical Density Based Clustering Leland McInnes and John Healy" year 2017 by Leland McInnes and John Healy.