Classification of Map Reduce

Neetu chhabra[1], Manish joshi[2]

[1]*Student, College of Computing Sciences And Information Technology,Teerthanker Mahaveer University, Moradabad*

[2]*Associate Professor, College of Computing Sciences And Information Technology,Teerthanker Mahaveer University, Moradabad*

[1]manish.computers@tmu.ac.in

[2]neetuchhabra0012@gmail.com

Abstract: **Map reduce implementation for processing and garneting big data sets. it is a programming model every task in the real world are expressible in this model as shown in paper it can automatically program or write in a functional style it execute on big cluster in machine of commodity it schedules the program and run time system be careful of portioning the data of input. it executes through set of machine management, machine failures.**

I. INTRODUCTION-

In this generation technology is increasing rapidly so that the size of data is also increasing. we are living in a world where data is boon now a days. the term big data came into existence due to the advanced technology. The big data is unable to store in a typical database because it has a dataset of huge size. Simple relational database management system cannot analyzed the huge dataset Generally the huge data can be store and process by the relational database management system but the data which is of huge amount can be structured, unstructured or semi structured

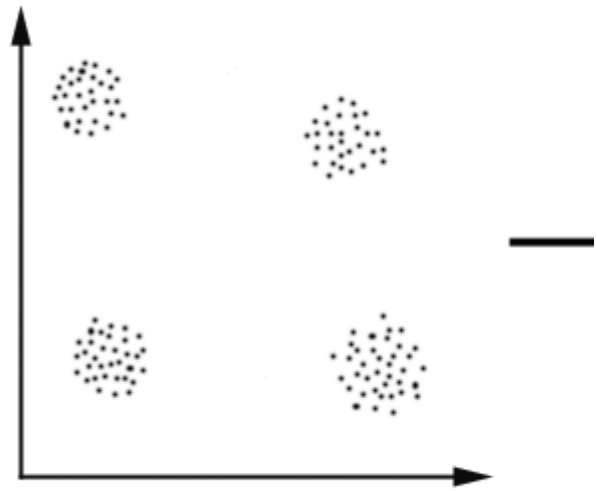
Researchers are worried with this continuously increasing of data dispensation which is storm of data is curving in almost all science research areas like web data, biomedical, Bio-Informatics and other disciplines due to its high accuracy and capability to deal with high dimension data researchers have the biggest challenge that how to do the analysis of the large scale of data so that they can get the meaningful result. to get better visualization of big data ,data mining comes into existence .To discover the new pattern from the dataset data mining is used. many researchers has used many algorithms of data mining there is a need to develop the algorithms of data mining which are suitable for analysing the big data .many other parallel algorithms have been developed but map reduce is practically suited for analysing the big data .in in this paper an algorithm for map reduce based clustering is used which run on several data sizes file and the time of training is calculated on hardtop cluster As there is a major problem with clustering that is to select the proper kernel parameters. This paper contains the number of several dataset by varying the value of penalty parameter C and RBF Kernel function parameter σ. Accuracy and Training time corresponding to them are same. as the paper is organised as follows

section 2 describe about the basics of clustering ,clustering advantages and disadvantage and why there is a need of clustering section three describe the architecture of parallel clustering section 4 defines the Hadoop framework which mainly focused on its two core components HDFS and map reduce distributed programming language.....section 5 includes the algorithm and again the architecture of Map reduce based on clustering Section6 includes the experimental results. And the last will include the future work.

II. 2. CLUSTERING:

Clustering could be calculated as the most important *unproven learning* problem; so, as every other problem of this kind, it deals with finding

a *arrangement* within a collection of unlabeled data.
Another definition of clustering could be "the method of manage objects into groups whose members are parallel in some way".
A *cluster* is therefore a collection of objects which are "similar between them and are "conflicting to the items belong to other clusters.
We can more elaborate our views via an example



In this sheath we make out the 4 clusters in which the information can be divided; the parallel criterion is *distance*: if the cluster belong to the two or more objects are close according to the distance. This is called *distance-based clustering*.

A further kind of clustering is *conceptual clustering*: two or more things belong to the same cluster if one defines a idea *familiar* to all that objects. In other prose stuff are grouped according to their fit to eloquent concepts, not according to simple match process

PARALLELAccording to the existing problem of the traditional means efficient parallel clustering algorithm (Par3PKM) - and that algorithm in processing massive traffic data. It can efficiently cluster a huge number of GPS trajectories of taxicabs the evaluation results demonstrate that the Par3PKM algorithm. In particular, it offer a practical reference for implementing that parallel computing of the same type .

After preprocessing of data , with the DTSAD method, we retrieve the relevant attributes (e.g., longitude, latitude) of the GPS trajectory records where the passengers pick up or drop off taxis from the abovementioned data sets. Then, on a Hadoop cluster with MapReduce, we cluster the retrieved trajectory data sets through the Par3PKM algorithm with , which is the number of own clusters.

HIERARCHIALHierarchical clustering, also tarnished as Connectivity based clustering, which is based on the inner concept of objects being more allied to nearby objects than to things beyond away. Hierarchical clustering is a method of cluster analysis that constructs the clusters or groups by recursively partitioning the instances in either a top-down or bottom-up approach. It build a cluster ladder or a ranking of cluster. . This cluster chain of command is known as a dendogram, which is a two dimensional map. Each cluster node consists of child clusters, sibling clusters division the points sheltered by their common parent. In this , each article is assigned to a cluster in such a way that if we have N items then we have N clusters. Here we find closest pair of clusters and merge them into single cluster. Distances are computed between new and old clusters..
The Hierarchical clustering algorithms can be subdivided into two types.
1) Agglomerative Hierarchical (Bottom-up): In this type of clustering, each object mainly exhibit a cluster of its own. Begin with "n" clusters and a single sample or

point indicates one cluster. Then the most similar clusters Ci and Cj are found and are merged into one cluster. Repeat step second until the number of cluster becomes one.

2) Divisive Hierarchical Clustering (top-down): This is a top down clustering technique in which all the objects or data points primarily belong to one cluster. Then the single cluster splits into two or more clusters that have high dissimilarity between them and this process continues until the desired cluster structure is obtained.

4. Hadoop Framework- It is an open-source software which encourage scattered application. It allow user function to be in touch and work with several autonomous computer nodes and terabytes or even petabytes of data. Goggle introduce a Google File System (GFS) and Google Map Reduce white papers in the year 2003 and 2004. The important feature of HF is division of the data into thousands of machines and carry out it in equal manner. The Hadoop cluster can be unit by plainly using article of skill hardwares. These servers can practice large scale data competency. The international Journalof Database of Theory and Application works with Hadoop Distributed File System (HDFS) and Map Reduce Distributed Programming Model.

The architecture of Hadoop framework is here:

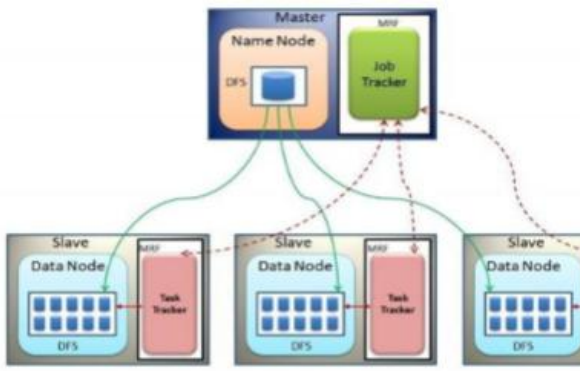


Figure . Hadoop Cluster

The HF consists of Hadoop common package containing all required JAR files to commence Hadoop.

Hadoop cluster is shown in the diagram.

The Hadoop Distributed File System (HDFS)

It is a scattered and scalable file system for Hadoop framework. HDFS is written in java and is a suitable file system of Hadoop. HDFS stores all its metadata to its devoted server known as NameNode also called master node. NameNode is the first node through which the user communicate to perform any input and output to the Hadoop cluster. There is only one master node in a Hadoop cluster and it should be the most reliable node of the whole cluster because without NameNode the whole cluster becomes unserviceable. It is the single point of failure of whole system. The actual data is stored in DataNodes also called slave nodes. HDFS is mainly designed for batch processing which provides high throughput and high I/O access of information. The Architecture of HDFS is shown below. The diagram clearly explains, the NameNode is responsible for storing metadata and the DataNode is responsible for storing the actual data.

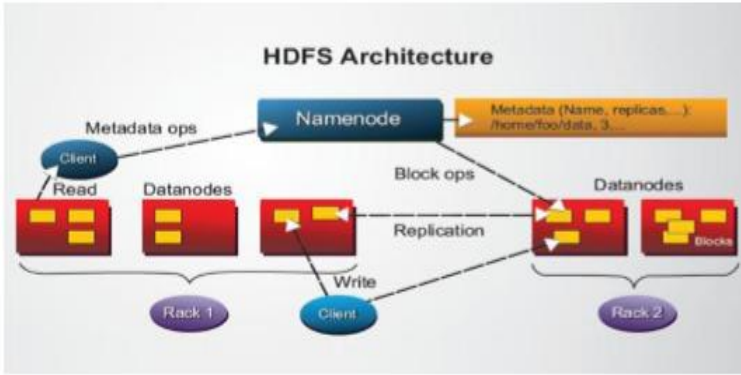


Figure 4. HDFS Architecture

Figure . HDFS Architecture

MapReduce Programming Model

It is a programming model introduce by Google in year 2004. MapReduce

programming model works on two functions called Map and Reduce. Users define a map function which is applied on input data in the form of key/value pair and generates a set of intermediate key/value pair. The reduce function combine these intermediate values corresponding to similar intermediate key.

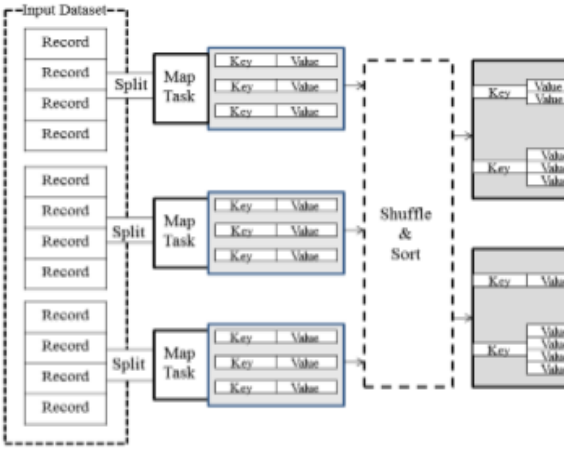


Figure. The MapReduce Programming Model

It shows how the input is divided into logical chunk and partitioned into various separate sets. These sets are then sorted and each sorted chunks are passed to the reducer. MapReduce model implements Mapper and Reducer interfaces to implement the map and reduce function.

III. CONCLUSION

Mining of data is still a big research area for large scaled data. Clustering is considered as the most effective classifier.. Most commonly used sequential Clustering is a very difficult task to work with large scale data set. Several experiments have been performed by many researchers .As it is proven that if we increased the data size and number of nodes on Hadoop cluster execution by the mean time it will decreases. By the Researchers it has been analyzed that a MapReduce based parallel SVM works efficiently on large datasets as compared to the sequential clustering

Advantage of using MapReduce based clustering over sequential clustering is the core components of Hadoop framework HDFS and MapReduce distributed programming model provides the data awareness between the NameNode and DataNode and also between the Job Tracker and Task Tracker.

IV. REFERENCES

[1] J. Lampi, ―Large-Scale Distributed Data Management and Processing Using R, Hadoop and MapReduce‖, University of Oulu, Department of Computer Science and Engineering. Master's Thesis, (2014).

[2] G C Fox, X H Qiu et al. Case Studies in Data Intensive Computing: Large Scale DNA Sequence Analysis. The Million Sequence Challenge and Biomedical Computing Technical Report, 2009

[3]Blake and C.J. Merz. UCI repository of machine learning databases, 1996. http://www.ics.uci.edu/mlearn/MLRepository.html

[4]M. Boull´ e. A grouping method for categorical attributes having very large number of values. In P. Perner and A. Imiya, editors, Proceedings of the Fourth International Conference on Machine Learning and Data Mining in Pattern Recognition, volume 3587 of LNAI, pages 228–242. Springer verlag, 2005b.

[5] Remzi H. Arpaci-Dusseau, Eric Anderson, Noah Treuhaft, David E. Culler, Joseph M. Hellerstein, David Patterson, and Kathy Yelick. Cluster I/O with River: Making the fast case common. In Proceedings of the Sixth Workshop on Input/Output in Parallel and Distributed Systems (IOPADS '99), pages 10.22, Atlanta, Georgia, May 1999.

[6] Arash Baratloo, Mehmet Karaul, Zvi Kedem, and Peter Wyckoff. Charlotte: Miscomputing on the web. In Proceedings of the 9th International Conference on Parallel and Distributed Computing Systems, 1996.

[7] Luiz A. Barroso, Jeffrey Dean, and Urs H¨olzle. Web search for a planet: The Google cluster architecture. IEEE Micro, 23(2):22.28, April 2003.

[8] John Bent, Douglas Thain, Andrea C.Arpaci-Dusseau, Remzi H. Arpaci-Dusseau, and Miron Livny. Explicit control in a batch-aware distributed _le system. In Proceedings of the 1st USENIX Symposium on Networked Systems Design and Implementation NSDI, March 2004.

[9] Guy E. Blelloch. Scans as primitive parallel operations. IEEE Transactions on Computers, C-38(11), November 1989.

[10] Armando Fox, Steven D. Gribble, Yatin Chawathe, Eric A. Brewer, and Paul Gauthier. Cluster-based scalable network services. In Proceedings of the 16th ACM Symposium on Operating System Principles, pages 78. 91, Saint-Malo, France, 1997