International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 04 | Apr -2017 www.irjet.net p-ISSN: 2395-0072

K Mean Clustering for Face book Data Set using Hadoop

Puja N. Vyavahare[1], Archana C. Bhavare[2], Vaishali Londhe[3], Harish Barapatre[4]

[1,2] M. E Student, Dept. of CE ,YTIET, Karjat , Maharashtra, India

[3]HOD, Dept. of CE,YTIET, Karjat, Maharashtra, India

[4]Ass. Prof. Dept. of CE, YTIET, Karjat, Maharashtra, India

Abstract – *Since 2004 Apache Hadoop gained popularity for parallel data processing. Hadoop is the concept of storing big data sets in distributed storage using computer cluster. Plenty of companies like Amazon , Google, Facebook , Yahoo etc. uses Hadoop as a parallel data processing architecture. Billions of users on social networking sites like face book, store or share their data. This paper studies architecture of Hadoop, K mean clustering algorithm which can be used for storing large data sets on face book using Hadoop.*

***Key Words*: K mean Clustering, Hadoop, Big Data Set , HDFS.**

1. INTRODUCTION

Social networking site Face book gained very much popularity in people worldwide. Everyday millions of users share their information in form of text, images or videos now days on their billion of pages. Facebook engineers or analysts manipulate this large data set using Hadoop. Data set is of 1 Peta byte disk space while 2500 CPU cores.[1]

Hadoop is an open source distributed framework which is used for distributed storage. It is founded by Apache foundation & usually processes large data sets. Hadoop has distributed file system. Large files distributed in small sets & referred as cluster. Packets transferred to the cluster nodes in parallel form. The programming model for big data is Map reduce programming model. Initially Hadoop referred Java platform for programming model. [2]

1.1 Hadoop for Facebook

Facebook used SQL query model for mapping & processing large data. Reason why facebook selected Hadoop is its high building reliability in each application. Because of large shared data, everyday plenty of nodes fail. As the cluster architecture is not fixed, it is difficult to count number of nodes in a cluster. Facebook engineers used Hadoop architecture for searching, log in, data warehouse, video & image analysis.[1,2]

1.2 Architecture of Hadoop

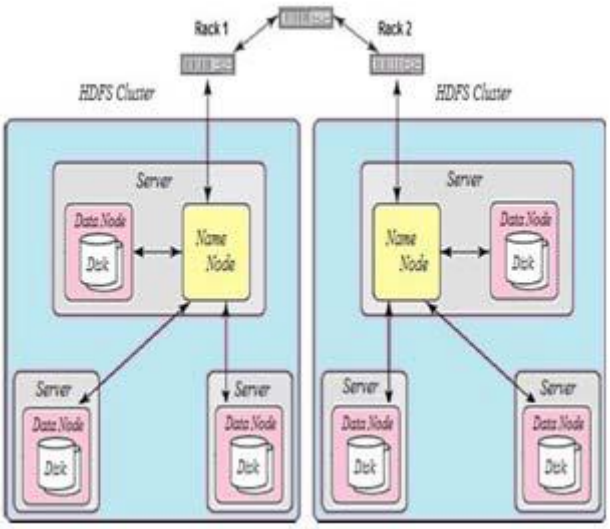

Fig. 1 Architecture of Hadoop

It is two level architecture visually Map Reduce layer & HDFS layer. Location of user can be tracked by using rack. Nodes are nothing but the commodity PCs. A rack consists of at least 30 to 40 nodes. A large data shared to every node replicate at every shared node by HDFS layer. HDFS is distributed file system where data found on distributed format. Replica produced on a local node while other two on the same rack & additional anywhere else in HDFS. This reduces failure of data at a node. A failed data node can also be detected.[3]

1.3 Hadoop Distributed file system (HDFS)

In HDFS every node assigned with a name. Hadoop has single name node while others are cluster of data nodes. Redundancy maintained on name nodes, data

nodes serves other block nodes. Entire cluster of nodes has single name space. Data coherency maintained by writing once & reading many times. [4]

Entire metadata is in main memory of server. Metadata is in form of list of files , files attributes , list of blocks & data nodes. A transaction log includes record of file creation, deletion.[5]

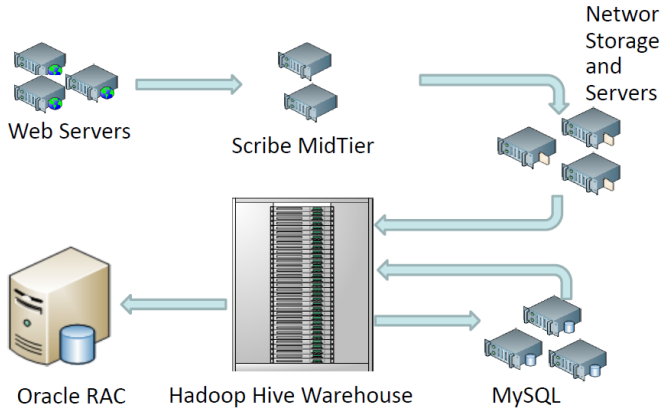

Fig.2 Hadoop cloud Data flow

1.4 Map Reduce Engine:-

It consists of job tracker & task tracker. Client application submits their job to job tracker. Job tracker pushes work to the task tracker node in cluster. The data passed to the nearest machines or known nodes in the cluster. Network traffic can be reduced by such priority based allocation in nodes. If time out condition occurs track tracer reschedule job. A heartbeat is sent from the Task Tracker to the Job Tracker every few minutes to check its status. Time delay & system load are not considered.[6]

Scheduling arranged on basis of FIFO schedule. Facebook developed fair share scheduler. Here scheduler is unaware of memory need for job. In real time applications in future a scheduler will analyze memory consumption also slots required to complete a job. Hence facebook has fast response.[7]

2. Proposed method (K Mean Clustering) :-

K Mean Cluster analyses the data in data mining. It partitions & observes K numbers of clusters data with respect to nearest mean which serves as a prototype to the cluster. The Lloyed's algorithm used for K mean clustering based on iterative refinement technique.

Algorithm works in fallowing 4 steps.

1. Random generation of data domain.
2. Creation of cluster by observation, nearest mean & partitioning.
3. Declaration of new mean by centroid of K cluster
4. Repeat step 2 & 3 until convergence.

Optimization & final results depends on initial cluster values. Time required for the conversions can be exponentially increases if number of cluster increases. The algorithm has polynomial smoothed running time. Algorithm has fallowing advantages.

1. It is applicable to univariate data.
2. It uses median instead of mean to normalize data.
3. Upper bound limit can be provable by K mean.
4. Suitable for text, image data sharing.
5. Triangle inequality speed up K mean clustering
6. Optimal number of clusters is possible by some changes in algorithm. Cluster size is deterministic.

3. CONCLUSION

A large attention gained by the K means clustering method for highly complex large data storage on facebook. Map reduce engine reduces network traffic by proper clustering. Cluster size should be declaration done by watching traffic or load on the network. Scalability as the size of the dataset increases by K mean clustering method. Hence the algorithm can be implemented in association with Hadoop.

REFERENCES

[1]https://www.facebook.com/notes/facebook-engineering/hadoop/16121578919/

[2] Apache Hadoop. http://hadoop.apache.org.

[3] Hadoop Distributed File System http://hadoop.apache.org/hdfs.

[4] Borthakur, D. 2007. The Hadoop Distributed File System: Architecture and Design. http://hadoop.apache.org/common/docs/r0.18.0/hdfs _ design.pdf.

[5] Shvachko, K., et al. 2010. The Hadoop Distributed File System.IEEE. http://storageconference.org/2010/Papers/MSST/S hva chko.pdf.

[6]Apache Hadoop. http: // en.wikipedia.org / wiki/ Apache_Hadoop.
[7] Joey Joblonski. Introduction to Hadoop. A Dell technical white paper. http://i.dell.com/sites/content/business/solutions/ white papers/en/Documents/hadoop introduction.pdf
[8] P.K. Agarwal and C.M. Procopiuc, "Exact and Approximation Algorithms for Clustering," Proc. Ninth Ann. ACM-SIAM Symp. Discrete Algorithms, pp. 658-667, Jan. 1998.
[9] K. Alsabti, S. Ranka, and V. Singh, "An Efficient k-means Clustering Algorithm," Proc. First Workshop High Performance Data Mining, Mar. 1998.
[10] S. Arya, D.M. Mount, N.S. Netanyahu, R. Silverman, and A.Y. Wu, "An Optimal Algorithm for Approximate Nearest Neighbor Searching," J. ACM, vol. 45, pp. 891-923, 1998