ISSN 2395-1621

Implementation of Kmeans Algorithm with MapReduce in Hadoop Distributed Environment

[#1]Mr. Mahesh Verma, [#2]Mr. Amol Waghole, [#3]Mr. Tejas Waghole,
[#4]Prof. Mr. S.R.Todmal

[1]mcoltd65@gmail.com
[2]amolwaghole42@gmail.com
[3]mandarn09@gmail.com

[#123]Department of Information Technology
[#4]Prof. Department of Information Technology

Jspm's
Imperial College Of Engineering & Research,
Wagholi, Pune-412207.

ABSTRACT

Data Mining and High Performance Computing are two broad fields in Computer Science. The k-Means Clustering is a very simple and popular data mining algorithm that has its application spread over a very broad spectrum. MapReduce is a programming style that is used for handling high volume data over a distributed computing environment. This paper proposes an improved and efficient method to implement the k-Means Clustering Technique using the MapReduce paradigm. The main idea is to introduce a combiner in the mapper function to decrease the amount of data to be written by the mapper and the amount of data to be read by the reducer which has considerably reduced the redundant MapReduce calls that have resulted in a significant reduction in the time required for clustering as it has decreased the read/write operations to a large extent. The implementation of Improved MapReduce k-Means Clustering has been clearly discussed and its effectiveness is compared to the regular implementation in an experimental analysis.

ARTICLE INFO

Article History

Received: 21st May 2016

Received in revised form :

22nd May 2016

Accepted: 25th May 2016

Published online :

28th May 2016

Keywords: **Data Mining, Clustering, K-Means Clustering, Distributed Computing, MapReduce, Hadoop**

I. INTRODUCTION

Data mining has been defined as the use of algorithms to extract information and patterns as part of Knowledge Discovery in Databases (KDD). The three important research areas of data mining are Classification, Clustering and Association rule mining. Brief reviews of these areas are discussed as follows. Classification assigns items to appropriate classes by using the attributes of each item. Clustering methods group items, but unlike classification, the groups are not predefined. A distance measure, such as the Euclidean distance between feature vectors of the items, is used to produce the clusters. Association rule mining (ARM) considers shopping-cart or market basket data items, i.e., the items purchased on a particular visit to the supermarket. ARM firstly finds out the frequent sets out of the data, which must have to meet a certain support level. Cluster analysis has been widely used in numerous applications, including data analysis, pattern recognition, market research, and image processing. In many businesses which involves sales and transactions, clustering can help marketing specialists discover distinct groups in their customer bases and characterize customer groups based on purchasing patterns. In biological field, it can be used to derive animal and plant taxonomies, categorize genes with similar functionality, and obtain an insight into structures inherent in populations. There are plenty of clustering algorithms available for use, among which K-Means clustering algorithm is one of the simple and popular algorithm. Kmeans is one of the simplest and popular unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a very simple and easy way to cluster a given data set through a certain number of clusters (assume k clusters) specified in prior. The main idea is to define k centroids, one for each cluster. In general, distributed computing is any technique of computing that involves multiple computers remote from each other that each has a role in a computation problem or information processing. ApacheTM Hadoop is one such open source framework that supports distributed computing. It came into existence from Google's MapReduce and Google File Systems projects. It was actually created by "Doug Cutting", the creator of Apache Lucene project which is

actually the widely used text search library. Hadoop finds its origin in Apache Nutch project, an open source web search engine, itself a part of the Lucene project. It is a framework that can be used for distributed computing and large-scale data processing, although it is best known for MapReduce and its distributed filesystem (HDFS). The Hadoop framework takes into account the node failures that can occur in its cluster and is automatically handled by it. This makes Hadoop really flexible and a fault tolerant platform for data intensive applications. This not only reduces the time required for completion of the operation or processing of voluminous data sets but also reduces the individual system requirements for computation of large volumes of data. Since the start of the Google File Systems(GFS) and MapReduce concepts, Hadoop has taken the world of distributed computing to a greater extent of successful track with various versions of Hadoop that are now being in existence and fewer under Research and Development. Few of which include Pig, Hive, Zookeeper, HBase, Sqoop, Oozie. The MapReduce structure gives great flexibility and speed to execute a process over a distributed Framework. Unstructured data analysis is one of the most challenging aspects of data mining that involve implementation of complex algorithms. The Hadoop Framework is designed to give the solution for the storage and computation of voluminous data sets. This can be done by downscaling and consequent integration of data and reducing the configuration demands of systems participating in processing such huge volumes of data. The workload is shared by all the nodes of computers connected on the network (Cluster) and hence increases the efficiency and overall performance of the network and at the same time facilitating the fast processing of voluminous data.

II. CLUSTER ANALYSIS

Data mining is truly multidisciplinary topic which can be defined in many different ways. There are a number of data mining functionalities are used to specify the kinds of patterns to be found in data mining task. These functionalities include characterizations and discrimination, the mining of frequent patterns, associations and correlations, classification and regression; clustering analysis, outlier analysis. Clustering is one of the most interesting topics in data mining. Clustering has its root in many application areas such as biology, image pattern recognition, security, business intelligence and Web search. The aim of clustering is to find intrinsic structures in data, and organize them into meaningful subgroups for further study and analysis. The basic concept of cluster analysis or clustering is the process of partitioning large data set of objects into small subsets. Each small subset is a unique cluster, such that the objects are clustered together based on the principle of maximizing the intraclass similarity and minimizing interclass similarity. Similarity and dissimilarity are assessed based on the attribute values describing objects and different distance measures. We measures object's similarity and dissimilarity by comparing objects with each other. These measures include distance measures like Euclidean distance, Manhattan distance, supremum distances between two objects of numeric data. Cluster analysis is a broad subject and hence there are abundant clustering algorithms available to group data sets. Very common methods of clustering involve computing distance, density and interval or a particular statistical distribution. Depending on the requirements and data sets we apply the appropriate clustering algorithm to extract data from them.

III. MATHEMATICAL KEYWORDS

$$M=\{ \Phi, \sum, \delta, q0, f \}$$

$$\Phi = \{A,B,C,D,E,F,G,H,I\}$$
$$\sum = \{P,KP,UP,BE,NP,AP,A,SA,C\}$$
$$q\,0 = A$$
$$f = I$$

Where

P = packet
KP = Known Packet
UP = Unknown Packet
BE = behavior Extraction
NP = Normal Packet
AP = Abnormal Packet
A = Alert for Intrusion
C = Clustering for Alert Aggregation

IV. PROPOSED SYSTEM

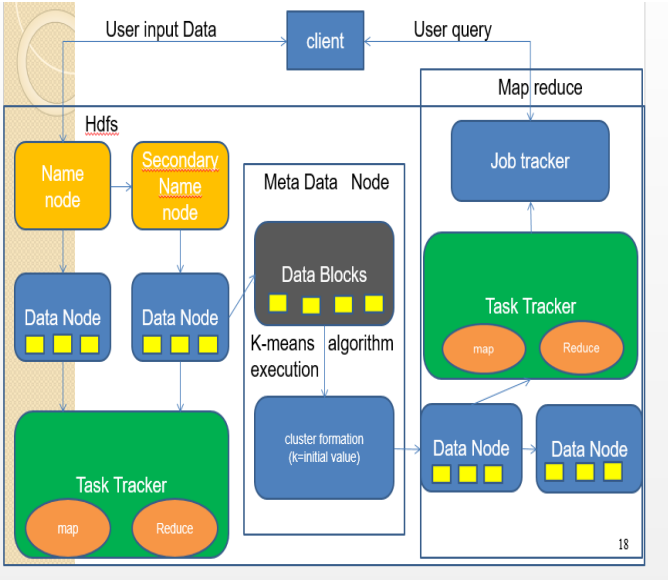

Fig 1. System architecture

The Proposed system will have better efficiency for processing small quantity of data sets. The Time Complexity of K-means algorithm for Data Processing will be less than of existed system. Faster Data Processing for smaller Data Sets and accurate Results for Better Analysis of Data.

K-means Algorithm

The K-means algorithm can be implemented as follows.

Designate k randomly selected points from n points as the centroids of the clusters.

Assign a point to the cluster, whose centroid is the closest to it, based on the Euclidean or some other measure;

Recompute the centroids for all the clusters based on the items assigned to them,

Repeat steps (2 through 3) with the new centroids until there is no change in point membership.

V. CONCLUSION

In this current world the data size is growing exponentially from various sources. It is very essential to process these huge data to extract useful information hidden in it. Clustering is one such research attention to extract useful information from voluminous data. Among several clustering algorithm K-Means algorithm is simple and popular. Hadoop is an apache open source product which gives us the distributive environment for processing of data. This paper discussed the implementation of K-Means Clustering Algorithm over a distributed environment in MapReduce programing model and improving the MapReduce process execution time.

REFERENCES

1. Implementation of k-means algorithm using hadoop framework. Uday kumar, etc.al, vol.3, Issue V, May 2015,ISSN 2321-9653.

2. Survey Paper On Clustering Techniques ,Amandeep Kaur Mann, etc.al , Vol.2,Issue 4,April 2013.

3. A Survey Of evolutionary clustering algorithms, Eduardo Raul Hruschka etc.al Vol.39 No.2, March 2009.

4.Survey of Recent clustering techniques in Data mining, Anoop Kumar Jain , etc.al, vol 1 Issue 1 Aug 2012 ISSN:2278-733X.

5. Dache : A Data Aware Caching for Big-Data Applications Using the Map-Reduce Framework , Yaxiong Zhao, etc.al, volume19,No.1 February 2014 ISSN 1007-0214 05/10 pp39-50.

6. https://hadoop.apache.org/docs/r1.2.1/single_node_setup .html

