IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-3, Issue-10 E-ISSN: 2347-2693

Generating Frequent Item Sets Using Apache Hadoop Map Reduce and Mahout

Barooru Sasanka Kasyap [1*] and K.Syama Sundara Rao [2]

[1*, 2] Department of CSE, Prasad V Potluri Siddhartha Institute of Technology, Vijayawada, A.P, India

www.ijcseonline.org

Received: Sep/25/2015 Revised: Oct/10/2015 Accepted: Oct /22/2015 Published: Oct /31/ 2015

***Abstract*—** The Item set Mining is one of the most well known techniques to extract knowledge from data. The mechanism having some problematic data, for those further enhancements have been applied based on the Big Data in which some performances has-been explores on Map Reduce machine. The new approach for mining large datasets such as K-means, Mahout which targets on speed and time while Big FIM is optimized to run on really large datasets. K-means algorithm depends on Map Reduce, which is the infrastructure for prepare more datasets of certain scattered clusters. These clusters are combined in the form of nodes and edges and also display the item sets. The execution of these clusters can also be implemented on large datasets also with high scalability and performance.

Keywords— Big Data;Hadoop; Map Reduce;Mahout;K-Means

I. INTRODUCTION

Data mining is the process to evaluate data from different aspect and incorporate the useful information which can be used to enlarge costs. Data mining is an incorporative process of complete arrangement in enormous statistics path at the intersection of systems. Large amount of data is taken from different viewpoints and constantly growing. Data storage has grown extremely from gigabytes to peta bytes and it cannot be fixed using popular database management system.

The frequent item set mining [3] is an amusing category of data mining that target on considering progression of events. Frequent item set mining, is the base data which takes form of sets of instances also called transactions which has a number of features called items. [3] Frequent item set mining can be used in several other ways for understanding the data and has been an essential part of data mining and also tries to extract information from database based on regular act.

Big Data point out to data sets whose range is above the capability of typical database software tools to represent the development material within an elapsed time. Big Data is a collection of large datasets that cannot be processed using traditional computing techniques rather it involves many areas of business and technology. Big Data capacity is uniformly affect object, as of 2012 extend against terabyte to peta byte of data. Big Data is used in economic sector largely in Face book over a large no of images are loaded and also in Google over 24 terabyte data is used every day. Big Data involves massive quantity, tremendous pace, and stretchable variation of data. The data can be in the form of structured and unstructured data.

Hadoop is an open source framework which allows for distributed and batch processing of large sets of data across cluster of commodity computers. [9] Hadoop is a java based programming framework that supports the processing of large amount of data in a distributed environment. The Hadoop ecosystem encloses Map Reduce along with HDFS. Hadoop Distributed File System that provides a high-throughput access to data. [6] HDFS is a scattered and extensible value-system for storing very large files reliably and to stream those data sets at high bandwidth to user application. An HDFS use a master/slave architecture in which master consists of single Name Node where as slave consists of one or more Data Nodes.

Map Reduce is a software framework easily to write applications which process the big data in parallel on large clusters. The two tasks are implemented on Map Reduce agenda such as Map Task and Reduce Task [5]. The key/value pair is used in map task and reduce task. In map task input data is converting into a dataset and disintegrate into key/value pair. After map task, the reduce task takes the output of map task as input and combine into a smaller tuples.

Mahout is a single which run on elephant as own master. The name comes from closest association with Apache Hadoop as an elephant as its logo Mahout offers the coder a ready-to-use framework for doing data mining tasks on vast extent of facts. Mahout lets applications to analyze large sets of data effectively and in quick time. Mahout provides Java libraries for common math's operations focused on linear algebra and statistics and primitive Java collections. Apache Mahout is a clear origin proposal i.e., mainly used for creating scalable machine learning algorithms.

II. RELATED WORK

K-means clustering act as strategy of vector quantization, originally from signal processing, i.e., famous as long as cluster investigation within data mining. K-means clustering aims to partitions [7] [8] n observations into k clusters in which each and everyone belongs to the cluster within the nearest mean, serving as a prototype of the cluster. K-means is specific about affecting freely research innovation so that to solve the well-known clustering problem. The procedure follows clean as well as accessible approach to rate agreed data set via simple and easy way to classify data file over un-mistakable of clusters assume k clusters in fixed apriority.[8] The reality take care of detail k centres,so as much as individual clusters main idea is to define k centers, one for each cluster. These centers should be placed in a way because of different section element contracting outcome, extremely powerful exceptional possibility is take care of field authority nearly manageable distant against each other.

In Data Mining, k-means uses the Euclidean distance [9] to find the distance between clusters. Euclidean distance is to find the distance between the clusters based on the dimensions position of the vectors in one dimension, the distance between two ends on the actual line is the entire desirability of their numerical difference. The distance between the two points (j, z) on the certain line, are given

by: $$\sqrt{[j-z]^2} = |j-z| \quad (1)$$

In this k-means we are taking the distance, for n - dimensional space, by using Euclidean distance is

$$d(j, z) = \sqrt{[j_1 - z_1]^2 + [j_2 - z_2]^2 + \cdots + [j_n - z_n]^2} \quad (2)$$

III. IMPLEMENTATION

Map Reduce is a functional programming used in Artificial Intelligence and it received highlight. Map Reduce is a substructure using for utilization to process huge amounts of data, in parallel, on large clusters of commodity hardware in a dependable aspects. [7]There are two tasks in map reduce program those are Map task and Reduce task. The important stages in map reduce programming are map per, combiner and reducer. In map per each item is split and counted. In combiner same items are combined and each count is given. Finally, in reducer each item counts are calculated and item with final count is displayed. Map, written by the user, takes an input pair and produces set of key/value pairs. Reduce, written by the user, accepts an intermediate key and set of values for that key [5]. Map (k1, v1) - > list (K2, v2) Reduce (K2, list (v2)) - > list (v2)). [5]

A. K-means clustering algorithm using Map Reducer:

K-Means Map Reduce algorithm runs on parallel map reduce framework such as Hadoop. The prune function is used to abolish the frequent items from a given data and clustered items. The algorithm starts with initializes [9] cluster items. In map step, the original problem is divided to sub-problems. The map tasks process smaller problems. Then they pass answers to reduce task. In reduce step, the reduce task collects answers of sub-problems and combine them to form the final answer of the original problem based on frequent item sets and clusters are calculated by joining, sorting and eliminating the equivalent nodes in map node [8] [9]

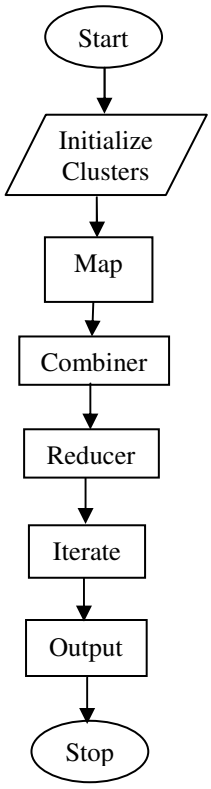


Figure-1: K-means clustering algorithm using Map Reducer.

The figure-1 shows us the rectangle (Map, Combiner, Reduce, Iterate, and Output) Circle shows the (start, stop) and arrow shows the flow line.
Steps for the flow chart for Map Reduce K-means Algorithm as shown in figure-1.

i. In this first cluster will be represented by their means are provided by user.

ii. In map step, each instance is mapped to a cluster by computing the nearest mean.
iii. In combine step, instances belonging to the same cluster are aggregated to one instance with high weight.
iv. In reducer step this instances in the same cluster are used to calculate cluster mean before that sort & shuffle will be done.
v. In iterate map, combine, and reduce steps until cluster means do not change.
vi. In output each instance's clustering label by nearest mean.

IV. RESULTS

Frequent items are mined from large data by K-Means. Appropriate input data is seized and apply K-Means program on that data as shown in figure -2:

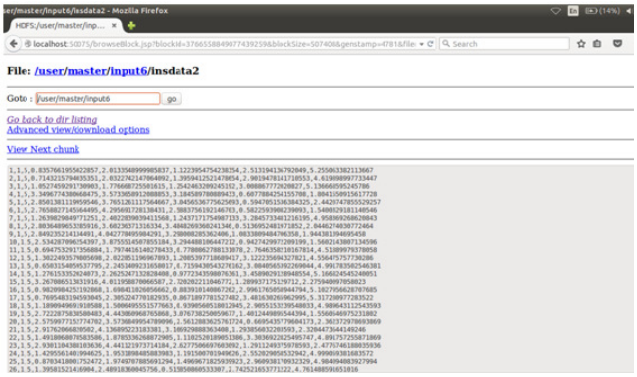


Figure-2: Input Data.

Now, appropriate clustered node data is given as shown in figure -3:

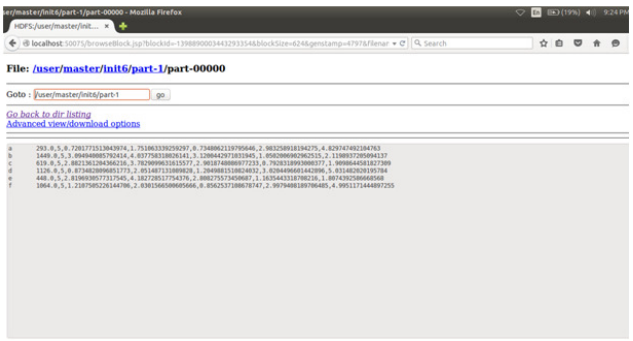


Figure-3: Clusters Data.

Finally, K-means program is applied to these input and clustered data. In K-means programming the input data is constant where as clustered are changed. The figure – 4: shows us the output is displayed in nodes as a, b, c, d, e, f along with clusters.

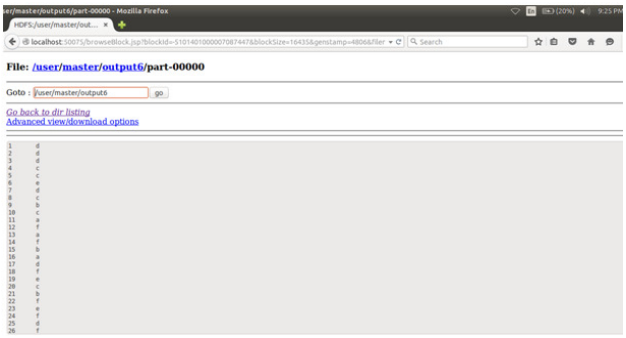


Figure-4: Output of K-means

In K-means the execution time for producing item set is 43sec. The figure -5: shows us the output of K-means program is seen by applying the input data and nodes data in the form of clustered graph with node id, label and also along with edges in form of colures.

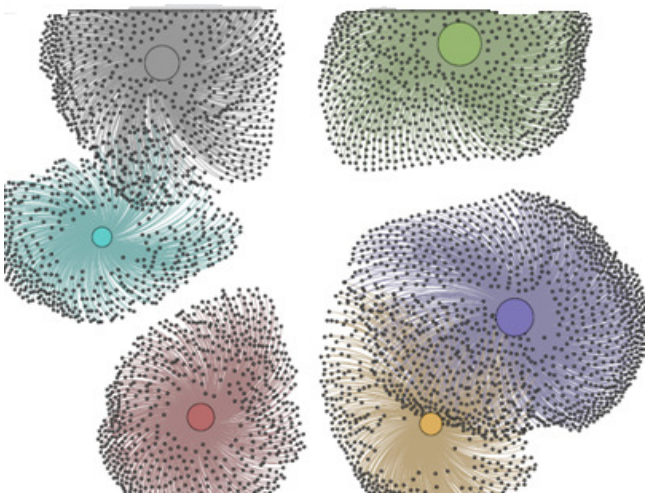


Figure-5: Output of K-means clusters using Gephi.

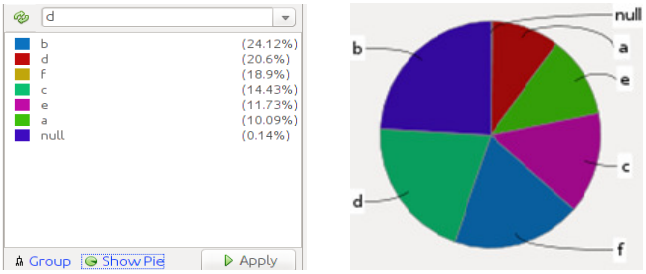


Figure-6: Output of K-means clusters using Gephi shows us the clusters in the form of "pie" graph.

The figure -6: shows us the output of K-means clusters using Gephi in the form of "pie" graph.

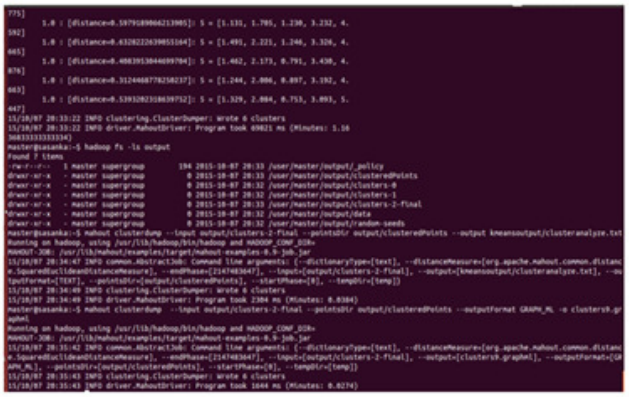


Figure-7: Output of Mahout

The figure-7: shows us the result of Mahout is seen in1min 22 sec.The result obtained in 43 seconds is better when compared with Mahout. In terms of execution time map reduce is better and In terms of Accuracy Mahout is best.

V. CONCLUSION

In K-means map reduce programming the clusters are generated in 43 seconds where as for Mahout the result is seen in1minute 22 sec.The result obtained in 43 seconds is better when compared with Mahout. In Map reduce we have to produce number of clusters information manually with specified parameter values. But In Apache Mahout It calculates the accurate number of clusters with automatic distance measure available in Mahout. So In terms of execution time map reduce is better and In terms of Accuracy Mahout is best.

REFERENCES

[1] Apache Hadoop project, http://hadoop.apache.org/

[2] Apache Mahout, http://mahout.apache.org/ , **2014**.

[3] Moens, S.; Aksehirli, E.; Goethals, B., "Frequent Item set Mining for Big Data," Big Data, **2013** IEEE International Conference Page No(**111, 118**), Oct. **2013**

[4] Srinath Parera, Thilina Gunarathane, "Hadoop Map Reduce Cook Book", [PACKT] publishing, ISBN: **9781849517287**, Page No. (**129-133**), Jan **2013**.

[5] Dean, Jeffrey, and Sanjay Ghemawat. "Map Reduce: simplified data processing on large clusters." Communications of the ACM volume 51, Issue 1, January **2008**, Page No (**107-113**). ISSN: **0001-0782** EISSN: **1557-7317.** In Proc.OSDI. USEXNIC, Association **2004.**

[6] Borthakur, D. "The Hadoop Distributed File System: Architecture and Design", **2007**

[7] W. Z. Zhao, H. F. Ma, Q. He. "Parallel k-means clustering based on Map Reduce". In CloudCom'09: Proceedings of the 1st International Conference on Cloud Computing, Page No (**674-679**), Berlin, Heidelberg, **2009**.

[8] Ping ZHOU, Jingsheng LEI, Wenjun YE, "Large-Scale Data Sets Clustering Based on Map Reduce and Hadoop", Journal of Computational Information Systems, **2011**

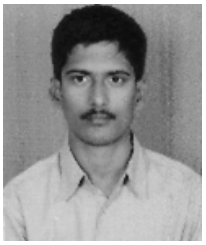
[9] Zhihua Li, Xugong Song, WenhuiZh, Yanxia Chen,"K-Means Clustering Optimizisation Algorithm Based on Map Reduce". ISCI. March **2015**.

[10] Jiawei Han and Michelin Kamber. "Data Mining, Concepts and Techniques". Morgan Kaufmann, **2001**

[11] Qing He, Fuzhen Zhuang, Jincheng Li, Zhongzhi Shi, "Parallel Implementation of Classification Algorithms Based on Map Reduce". Page No (**655-662**). 5th International Conference, RSKT, **2010**.

AUTHORS PROFILE

Mr. B. Sasanka Kasyap Received his B.Tech degree in



Computer Science & Engineering from Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh. And he is currently pursuing M.Tech Degree in Computer Science & Engineering in Prasad V. Potluri Siddhartha Institute of Technology (Autonomous) Vijayawada, Andhra Pradesh, India and is affiliated to Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh.

Mr. K.Syama Sundara Rao is presently, Assistant

Professor, Dept.of Computer Science & Engineering. Prasad V Potluri Siddhartha Institute of Technology (Autonomous) Kanuru, Vijayawada, A.P, India. He had completed B.Tech (CSE) & M.Tech (Web Technology). He had teaching experience more than 9years 3months. His areas of interests are Web Technology, Big Data. He is a life member of ISTE and CSI.