Parallel Analytical Model for Frequent Itemset Mining

Poorva K
Information Science and Engineering
VVCE, Mysuru

Anushree H K
Information Science and Engineering
VVCE, Mysuru

Mahesha K V
Information Science and Engineering
VVCE, Mysuru

Pavithra T R
Information Science and Engineering
VVCE, Mysuru

Vinutha D C
Assistant Professor
Information Science and Engineering
VVCE, Mysuru

Chandini S B
Assistant Professor
Information Science and Engineering
VVCE, Mysuru

***Abstract*— The data generated in real world applications in recent days are quite huge and processing and analyzing those data has become a challenging task. For any analytical and inference engine finding frequent itemsets turns out to be a major functionality. Frequent itemset mining is performed usually with the help of association rule mining technique in data mining. Generally the results obtain from these techniques are large and diverse which makes it difficult for an inference engine to reach a conclusion. One of the ideal solutions for handling such datasets is by devising a parallel processing system to efficiently run the data mining approaches and obtain a more accurate and easily analyzable output. The research carried out here aims at designing and developing parallel analytical model for frequent itemset mining in big data by integrating R on Hadoop.**

Keywords—Association rule mining, parallel mining algorithms, Inference engine, Hadoop

I INTRODUCTION

Big data: everybody is by all accounts discussing it, however what is big data truly? How is it changing the path scientists at organizations, non-benefits, governments, foundations, and different associations are finding out about their general surroundings? Where is this information originating from, how is it being prepared, and how are the outcomes being utilized? There is no hard and fast rule about precisely what estimate a database should be all together for the data within it to be viewed as "big." Instead, what ordinarily characterizes big data is the requirement for new methods and instruments so as to have the capacity to process it. With a specific end goal to utilize big data, you require programs which traverse numerous physical as well as virtual machines cooperating in show so as to process the greater part of the data in a sensible traverse of time. Getting programs on different machines to cooperate in a proficient way, so that each program knows which segments of the data to process, and afterward having the capacity to put the outcomes from the majority of the machines together to comprehend an expansive pool of data takes unique programming procedures. Since it is normally substantially speedier for projects to get to data put away locally rather than over a system, the circulation of data over a bunch and how those machines are arranged together are additionally vital contemplation which must be made when pondering big data issues.

Apache Hadoop was destined to improve the use and understand significant issues of big data. The web media was producing heaps of data regularly, and it was ending up noticeably extremely hard dealing with the data of around one billion pages of substance. All together of progressive, Google created another philosophy of preparing data prevalently known as MapReduce. Later following a year Google distributed a white paper of Map Reducing system where Doug Cutting and Mike Cafarella, motivated by the white paper and accordingly made Hadoop to apply these ideas to an open-source programming structure which bolstered the Nutch web index extend. Considering the first contextual investigation, Hadoop was outlined with considerably easier capacity framework offices. Apache Hadoop is the most critical structure for working with Big Data. Hadoop greatest quality is adaptability. It upgrades from working on a single node to thousands of nodes without any issue in a seamless manner.

Association rule mining is a methodology which is intended to discover visit designs, connections, associations, or causal structures from informational collections found in different sorts of databases, for example, social databases, value-based databases, and different types of data storehouses.

II. LITERATURE SURVEY

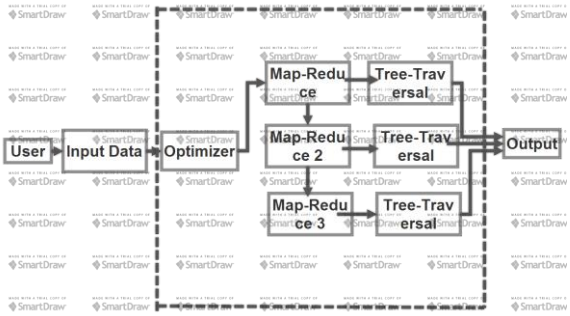
FiDoop: Parallel Mining of Frequent Itemsets Using MapReduce: Yaling Xun, Jifu Zhang, and Xiao Qin, *Senior Member, IEEE(2015).*

FIUT: A new method for mining frequent itemsets Yuh-Jiuan Tsay , Tain-Jung Hsu , Jing-Rung Yu (2009).

Comparing dataset characteristics that favor the apriori, eclat and FP-Growth frequent itemset mining algorithm :Jeff Heaton(2016).

III. PROPOSED SYSTEM

The review of existing literature portrays the various issues and performance challenges that exist in the current world big data analytics. The proposed system addressees the issues in frequent itemset mining that arises from large data analytics. The research proposes a parallel algorithm based platform for big data analytics which deals with efficient and accurate processing and analyzing of big data, here initially the data that is taken as an input from the user is fed into a prepossessing optimizer which discretize the data and takes it into a multi stage parallel algorithm engine. For running this parallel algorithm engine a Hadoop based platform is set up. The Hadoop gives the sufficient processing capabilities for the efficient execution of parallel algorithm. The parallel algorithms use some association rule mining algorithms which can be combined and parallely executed with multiple MapReduce approaches which does the improvement in accuracy and generate a very easily analyzable data. Each of the MapReduce performs operations on the same list of dataset which is handled with multiple algorithms to give the minimized output ,thus the system overcomes the performance issues as well as issues ariesed due to dimensionality of the dataset efficiently.



IV. EXPERIMENTAL RESULTS

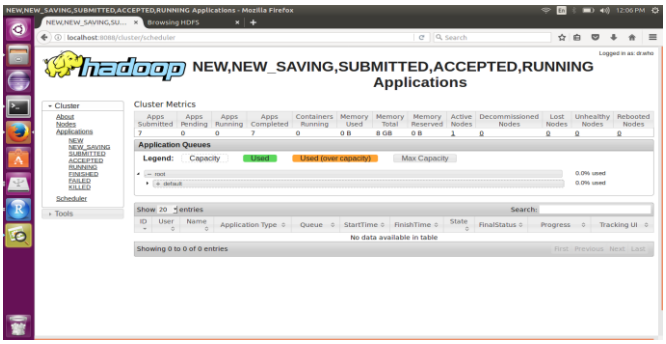


Fig 1.initalization of hadoop

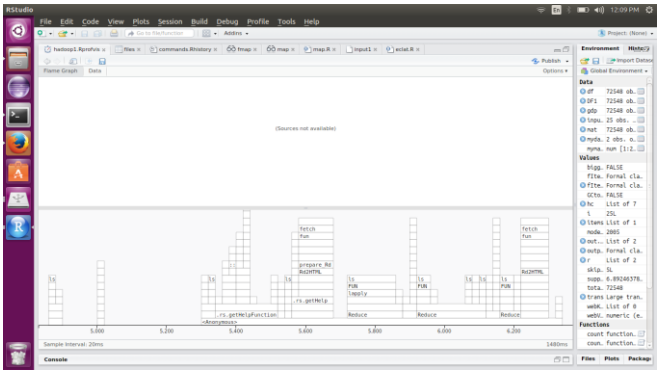


Fig 2.flame graph

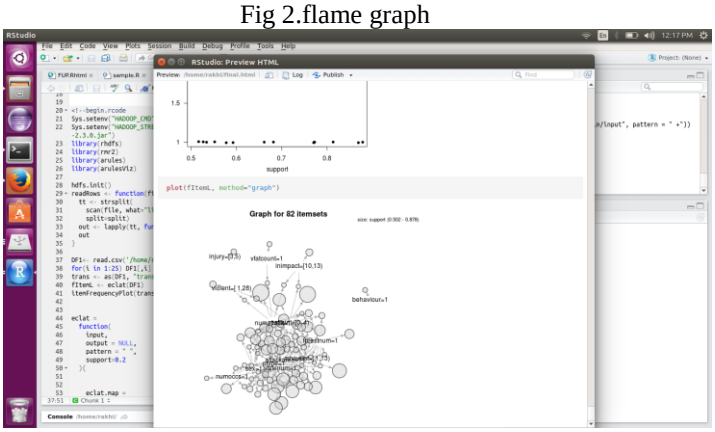


Fig 3.visualization of frequent itemset obtained

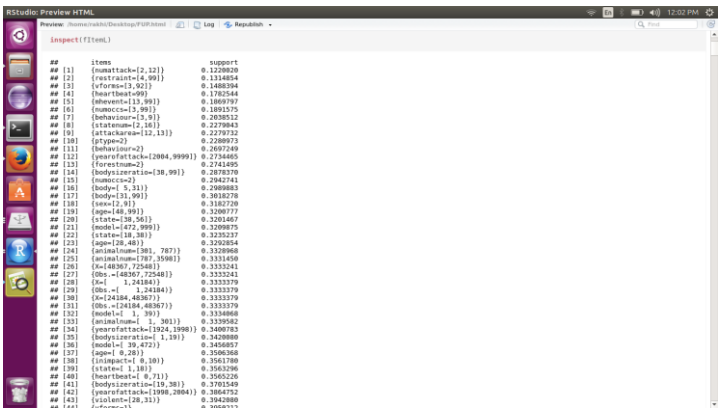
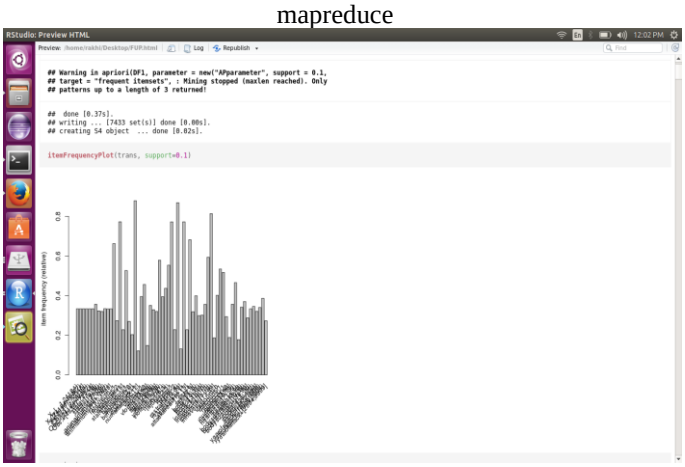


Fig 4.associate rule mining algorithm with multiple mapreduce



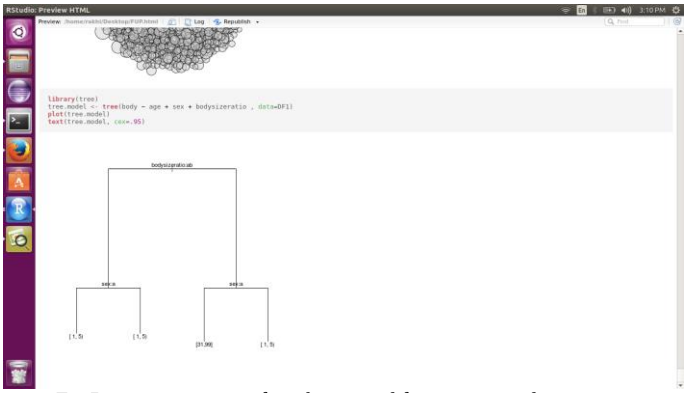


Fig 5 tree represtation of attributes used for associate rule mining

CONCLUSION

To run analytical processing parallelly in order to achieve efficient execution various algorithms are used. The parallel efficient model with the help of Hadoop and multiple parallel algorithms supporting ARM are executed in the platform and as a result obtaining accurate and easily analyzable frequent patterns from Big data, Association rule mining is the data mining technique used to analyze efficiency of the different algorithms

REFERENCES

[1] S. Kotsiantis and D. Kanellopoulos, "Association rules mining: A recent overview," *GESTS Int. Trans. Comput. Sci. Eng.*, vol. 32, no. 1,pp. 71–82, 2006.

[2] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach," *Data Min.Knowl. Disc.*, vol. 8, no. 1, pp. 53–87, 2004.

[3] J. Xie and X. Qin, "The 19th heterogeneity in computing workshop (HCW 2010)," in *Proc. IEEE Int. Symp. Parallel Distrib. Process.Workshops Ph.D. Forum (IPDPSW)*, Atlanta, GA, USA, Apr. 2010,

[4] J. Han, H. Cheng, D. Xin, X. Yan, "Frequent pattern mining: Current status and future directions", *Data Mining Knowledge Discovery*, vol. 15, no. 1, pp. 55-86, Aug. 2007.

[5] Farah Khan, Divakar Singh, "Knowledge Discovery on Agricultural Dataset Using Association Rule Mining", *International Journal of Emerging Technology and Advanced Engineering*, pp. 925-930, 2014.