IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-15, May 2019 E-ISSN: 2347-2693

A Novel Data AggregationTechnique for Removing Redundant Data in Hadoop

Uday Shankar S V[1*], AnveshNaik[2], Manoj C K[3], Praveen B[4], Yadush B R[5]

[1,2,3,4,5]Department of ISE, SJB Institute of Technology, India

DOI: https://doi.org/10.26438/ijcse/v7si15.270271 | Available online at: www.ijcseonline.org

***Abstract*:** Hadoop is the software framework which was developed by Apache Software Foundation.Hadoop framework is written in java with purpose to handle large amount of data. Hadoop manages huge volume of data.Hadoop runs the task under the MapReduce algorithm. MapReduce is a programming model suitable for processing of huge data. MapReduce framework has two phase, map phase and reduce phase.a mapredce job is usually splits the input data set into independent chunks,which is done by map phase.the framework sorts the output of the map which are input to reduce framework. To running frequent itemset require more resource and time consuming. To overcome this problem here we implementing the nobel data aggregation technique.

***Keywords*-** herewe are grouping the frequent itemsetand remove the redundant data.

I. INTRODUCTION

Big data is a term that describes the large volume of data-both structured and unstructureddata. The importance of big data doesn't revolve around how much data you have, but what you do with it. You can take data from any source and analyze it to find answer that enables cost reduction, time reduction, new product development and optimized offerings, and smart decision making.Traditional parallel frequent itemset mining techniques are focused on load balancing; data are equally partitioned and distributed among computing nodes of cluster. more often than not, lack of analysis of correlation among data leads to poor data locality. The absence of data collection increases the data shuffling costs and the network overhead, reducing the effectiveness of data partitioning. in this study, we shows the nobel data aggregation techniques to overcome the drawback of the traditional parallel frequent itemset mining techniques.

Data aggregation is a techniques in which, there are two steps are there data grouping and data aggregation. Data grouping identifies one or more data group based on values in selected features and then, data aggregation put together the values in one or more selected Colum for single group .by applying this technique, change the scale of data, and provides the data reduction and more stable data.

II. RELATED WORK

By embedding MapReduce with Hadoop is very much useful to collect huge amount of climatic data and to store various of parameters such as Humidity,Temprature etc on multiple nodes and clusters. MapReduce is nothing but a model for processing a task by splitting a huge task into serval others sub-task and it is distributed to different nodes to process in parallel. Hadoop helps to build a platform which is flexible and scalable to analyze extremely huge data nearly Pentabyes(10^{15}) of data. Amazon uses Hadoop to search their products and to process their million of sessions. Hadoop is a distributed file system due to this feature adobe uses for internal storage and processing.A part from these IBM, Rackspace, The New York Times, University of feriburg and lot more using Hadoop.

III. PROPOSED WORK

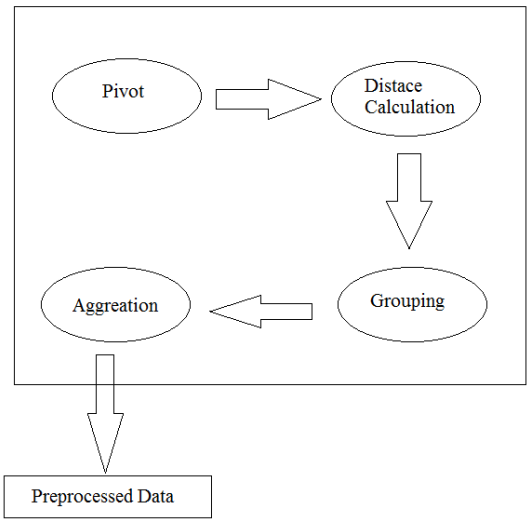

Fig 1. Data Preprocessing System

The above figure shows the propose*d preprocessing system of our project.here we, first randomly generate the pivot value, then we will compare this pivot value with centroid

value to find any match exit or not. Then we find the distance between the matching values by below formula.

$$\sqrt{(x2-x1)^2+(y2-y1)^2}$$

Using this formula we generate a list of values, and then we compare this value with pivot value. If the values are same, we group the matched values. Then we apply sorting and shuffling techniques. Finally we produce the preprocessed data.

IV. RESULTS

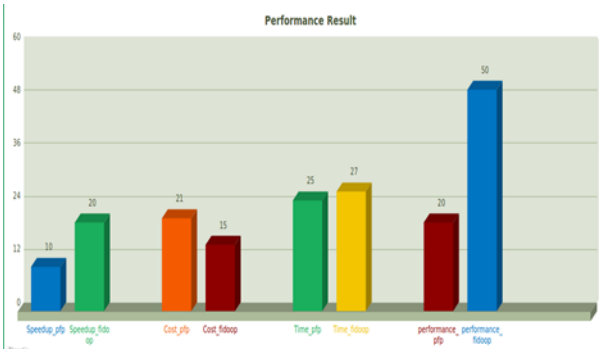

Fig 2. Bar Graph of Computation Time

The above graph shows the comparions between performance of the Hadoop. Using previous hadoop technique time taken to processdata is more and redundent data is high. By using data preprocessing technique computation time can be reduced and redundant data can be reduced as shown in the graph.

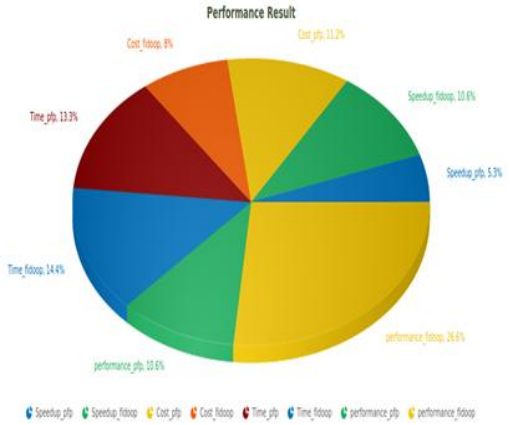

Fig 3. Pie chart of Performence

V. CONCLUSION

In previous techniques, to process millions of records generated by internet, smart device etc it was too time killing process. To overcome previous drawbacks we use MapReduce framework. Where it takes huge amount of datasets as inputs and splits the data and distributes across multiple nodes. so that processing can be done very quickly and reduces amount of time taken for processing. Using MapReduce with Hadoop data can be analyzed efficiently and redundant data can be reduced.

REFERENCES

[1]. Y. Xun, J. Zhang, and X. Qin, "Fidoop: Parallel mining of frequent itemsets using mapreduce," IEEE Transactions on Systems,Man ,and Cybernetics: Systems, doi: 10.1109/TSMC.2015.2437327, 2015.

[2]. J. Leskovec, A. Rajaraman, and J. D. Ullman, Mining of massive datasets. Cambridge University Press, 2014.

[3]. M. Liroz-Gistau, R. Akbarinia, D. Agrawal, E. Pacitti, and P. Valduriez,"Data partitioning for minimizing transferred data in mapreduce," in Data Management in Cloud, Grid and P2P Systems. Springer,2013.

[4]. T. Kirsten, L. Kolb, M. Hartung, A. Groß, H. K¨opcke, and E. Rahm,"Data partitioning for parallel entity matching," Proceedings of theVLDB Endowment, vol. 3, no. 2, 2010.

