 ISSN: 2454-4248
84 – 86

Sorting Technique- An Efficient Approach for Data Mining

Jyoti Karlupia
Computer Science & Engineering
Golden College of Engineering and Technology
Gurdaspur, India

Rohit Mahajan
Computer Science & Engineering
Golden College of Engineering and Technology
Gurdaspur, India

Mohit Angurala
Asst. Prof.
GNDU Regional Campus,
Gurdaspur,
Punjab, India

Abstract— As the new data or updates are arriving constantly, it becomes very difficult to handle data in an efficient manner. Moreover, if data is not refreshed it will soon become of no use. Hence data should be updated on regular mode so that it donot obsolete in coming future. In traditional work several other approaches or methods like page ranking, i2mapreduce(that is extension of MapReduce) were used to enhance performance and increase computation speed as well as run-time processing. But as we have seen the performance is not upto that level which is required in current environment. So, to overcome these drawbacks, in this paper sorting technique is proposed that can enhance mean value and overall performance.

Keywords—*Hadoop,single cluster,sorting*

*****

I. INTRODUCTION

As the technology is tremendously getting advanced day by day due to increase in use of social networking sites, therefore the amount of data is proportionally increasing. Hence the the amount of data that is coming out constitutes about ninety percentage of unstructured data. This is where hadoop comes into picture. Hadoop can handle any kind of whether structured or unstructured data in an efficient way like no other can. Hadoop is a framework which is open source and it is in java language. This framework can work in an distributed kind of environment where it also serves distributed storage function. There are ample of advantages of Hadoop which is shown in below figure:

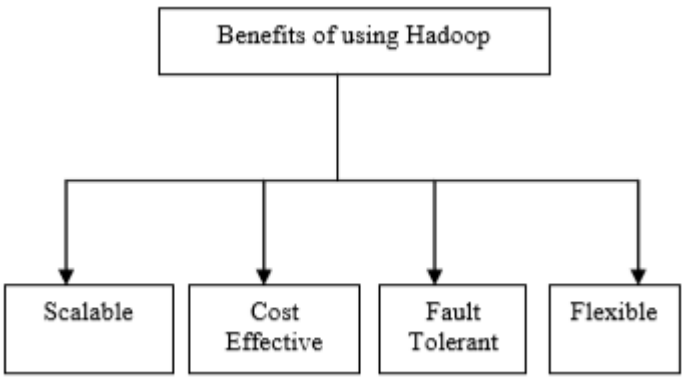

Figure 1: Advantages of Hadoop

- Scalable– Scalable means to can stretch it any time. In this scalable feature nodes can be easily attached to the existing nodes without modifying the existing formats.
- Cost effective– Parallel computing is implemented using hadoop. Hence it will help to reduce the cost per byte of storage. Therfore it can handle every sort of data.
- Flexible– As we have discussed in our introduction beginning that hadoop has a feature to adapt any kind of structured or unstructured data. In this case the data from different parts gets integrated and joined, therefore making it more flexible and better.
- Fault tolerant– Fault tolerant is another feature of hadoop in which if one node dies or fails the control will get redirected to some other node which inturn will not halt the work.

II. LITERATURE REVIEW

Yanfeng Zhang, Shimin Chen described a MapReduce-based framework for incremental big data processing. i2 MapReduce combines a fine-grain incremental engine, a general-purpose iterativemodel, and a set of effective techniques for incremental iterative computation. When implemented in a tool it shows that i2MapReduce can significantly minimizes the run time for refreshing big data mining results compared to re-computation on both plain and iterative MapReduce.

Harish, Kavitha: In this paper based on the architecture of Hadoop Distributed File System and Hadoop MapReduce framework present the methodology used in pre-processing of huge volume of web log files and finding the statics of website and learning the user behaviour.

Jie Song, Chaopeng Guo, Zhi Wang: This paper presents HaoLap (Hadoop based oLap), an OLAP (OLAP (Online Analytical Processing) system for big data. The paper illustrates the key techniques of HaoLap including system architecture, dimension definition, data storage, OLAP and data loading algorithm. The experiment results show that HaoLap boost the efficiency of data loading, and has a great advantage in the OLAP performance of the data set size and query complexity.

Paolo Nesi, Gianni Pantaleo, Gianmarco Sanesi: This paper presents a distributed framework for crawling web documents and

 ISSN: 2454-4248
84 – 86

running Natural Language Processing tasks in a parallel fashion. The system is based on the Apache Hadoop ecosystem and its parallel programming paradigm, called MapReduce. In the specific, they implemented a MapReduce adaptation of a GATE application and framework (a widely used open source tool for text engineering and NLP).

Bo Dong, Qinghua Zheng: In this paper, the relationship between file size and HDFS Write/Read (denoted as W/R for short) through-put, i.e., the average flow rate of a HDFS W/R operation, is studied to build HDFS performance models from a systematic view. These analysis results can provide effective guidance and implications for the design and configuration of HDFS and Hadoop-based applications.

III. PROBLEM FORMULATION

New data and updates are adding day by day, so to improve data mining application is must and incremental processing is a sophisticated approach to enhance mining result. They have proposed i2mapreduce i.e. incremental processing extension to MapReduce and widely used framework for mining big data. They have used fine grain incremental processing. Performing key value pair level incremental processing is better than task level computation to improve mean value. The traditional work has used four algorithm and i2map reduce technique to improve performance of hadoop and calculated the mean value of Hadoop. This mean value will actually show or prove performance of hadoop. Actually, the improvement is necessary because they did not sorted jobs to slaves and mean value was not good . we will improve mean value and sort the jobs to slaves and mean value will be better than the previous work. We will implemented a new technique that is sorting technique to further enhance the performance and the Mean value. Firstly we will create hadoop cluster and machines. Then we will connect with master slaves and then we will calculate mean value using machines.

IV. PROPOSED WORK

The development of the proposed algorithm has been divided into 2 phases:

PHASE -1

- Comprises of setting up a single node Hadoop cluster and gaining hands on hadoop .
- Coding the basic algorithm is also included in phase 1.

PHASE- 2

- Coding of the sorting algorithm and implementing the .jar file of the same on a single node hadoop in phase 2.
- The comparative analysis of the algorithms on the basis of mean values will also be done in the same phase .

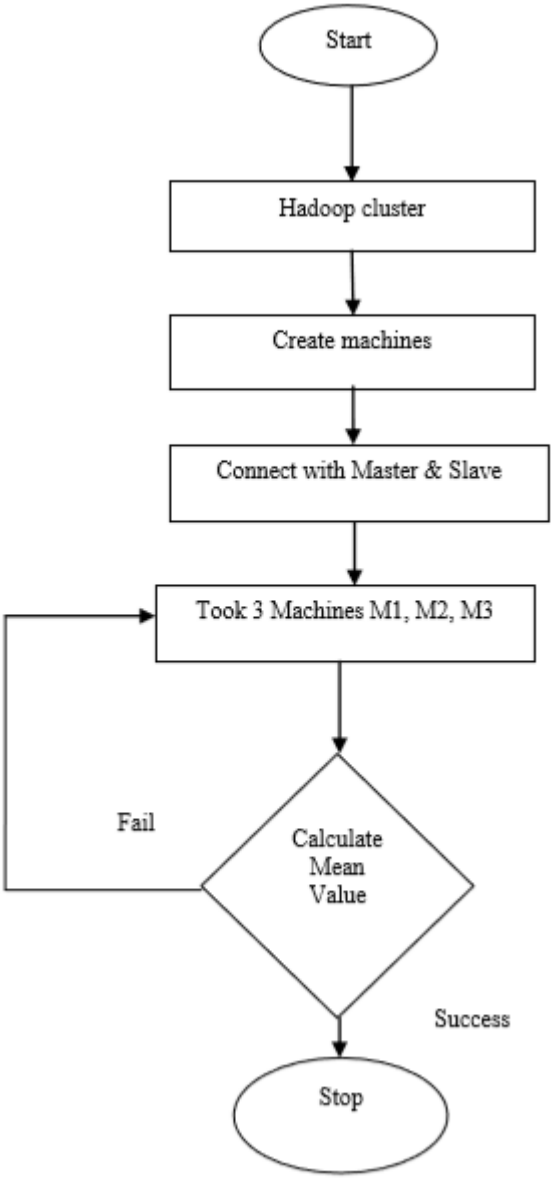

Fig 2. Proposed Flowchart for scenario

V. CONCLUSION

The previous work provide solution for data mining and computation with i2mapreduce and used fine grain incremental processing and minimize the problem of recomputation. Moreover, it also provides approach which only modify MapReduce using algorithm but the mean value was not upto the level. In this paper in order to increase the performance sorting technique is shown in the flowchart and this technique will definitely give better results when implemented.

REFERENCES

[1]. Yanfeng Zhang, Shimin Chen, Qiang Wang, and Ge Yu, Member, IEEE , “i2MapReduce: Incremental MapReduce for Mining Evolving Big Data” by in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 7, JULY 2015

 ISSN: 2454-4248
84 – 86

[2]. Harish S1, Kavitha G2 "STATISTICAL ANALYSIS OF WEB SERVER LOGS USING APACHE HIVE IN HADOOP FRAMEWORK" vol. 3, issue 5, may 2015.

[3]. Jie Song Chaopeng Guo Zhi Wang Yichan Zhang GeYu Jean-Marc Pierson "A Hadoop based OLAP System for Big Data" JSS 9385.

[4]. Paolo Nesi, Gianni Pantaleo, Gianmarco Sanesi "A Hadoop Based Platform for Natural Language Processing of Web Pages and Documents" YJVLC729.

[5]. Bo Dong, Qinghua Zheng, Feng Tian, Kuo-Ming Chao, Nick Godwin,Tian Ma, Haipeng Xu,"Performance models and dynamic characteristics analysis for HDFSwrite and read operations" 93 (2014) 132–151

[6]. Chen-Hau wang,Ching-Tsorng Tsai(2014) "A hadoop based Weblog analysis system" DOI 10.1109/U-MEDIA.2014.9.

[7]. Mr. Swapnil A. Kale1, Prof. Sangram S.Dandge2" UNDERSTANDING THE BIG DATA PROBLEMS AND THEIR SOLUTIONS USING HADOOP AND MAP-REDUCE" Volume 3, Issue 3, March 2014, ISSN 2319 – 4847.

[8]. Seyyed Mojtaba Banaei1, Hossein Kardan Moghaddam2*" Hadoop and Its Role in Modern Image Processing" 2014, 4, 239-245 2014, 4, 239-245

[9]. A Kavitha, S, Suseela, G, G Kapilya (2013),"Big Data", CSI Communication, ISSN 0970-647X,Issue No.1 vol.37, pp. 19.

[10]. Dr. Milind Bhandarkar (2013), " Big Data Systems: Past, Present & (possibly) Future " CSI Communication, ISSN 0970-647X, Issue No.1 vol.37, pp. 7-8.