A Heuristic Based Approach for the Big Data Processing Over Multiple Nodes

Aishwarya Dubey, Pranjal Tiwari, Pankaj Kumar Sahu
Dept. Of CSE, Gargi Institute of Science & Technology
Bhopal, India
Dubey.aish07@gmail.com
Pankajkumarsahu99@gmail.com

Abstract-Big-data provides techniques to process very large scale datasets. Big data can be a structured unstructured, semi-structured that depends on the data management methods. Big data is the data which presents in large scale form and required an efficient technique to process that data. Scheduling algorithms provides a way to process big data among the multiple sources. A new scheduling and ranking based technique is proposed in this paper which provides an enhanced functionality to process big data. That technique provides a mechanism to provide an optimization technique for the ranking and availability to check performance of the technique over the factors. A comparison analysis of the results for the existing technique and proposed technique presented in result and analysis section to provide an enhanced functionality to the user and efficient technique to process big data.

Keywords: **- Big Data, Scheduling, Ranking, Heuristics Technique.**

I INTRODUCTION

Big data is a term which contains large volume of data in both structured and unstructured format. That bury the business on day to day basis. But that data introduced by the various heterogeneous sources. Which scattered in random manner, traditional techniques are not suitable to process that data. Efficient and enhanced technique is required to process that data. Big data poses following features: velocity, data stream in big data contains very high speed, files like torrent process in near real time speed. Volume, data collected from various sources like social media, business transactions, or some other sources. To manage data in such large amount technologies like Hadoop are used. Data introduced by the various sources that generate inconsistency in the data.

Various scheduling techniques are used to provide a managed access of that data. But traditional techniques are not efficient to provide better performance to provide better access for that data. FIFO (First in First out) is a technique which used to provide Big Data processing mechanism in First In First Out manner but that technique is not able to provide a better scheduling mechanism in a large scale data scenario. Fair scheduling in that scheduling technique, in that technique fair amount of cluster space is provide to the every user in the system, like in Facebook each user can have their own cluster space can access resources simultaneously. Capacity based scheduling, in that type of scheduling, scheduling can be performed on the basis of the capacity of the system. But these techniques also suffers some defects. A new scheduling technique is required to provide better performance to process Big Data.

Figure 1.1: Big Data.

The Scheduling algorithm is a simulated evolutionary algorithm, is proposed for solving combinative expansion problems. In Scheduling algorithm is exist independent and parallel evolution. The competitions between tribes and family are used to increase the search efficiency. All clan are ranked a Scheduling based on their main function and allocated different search spaces according to their position in the line-up, which is favourable to balance local and global search. The Scheduling algorithm is applied to the solution of knapsack problem and the optimal design of pressure relief header network. Computational results reveal that the Scheduling algorithm is able to find optimal or near-optimal solutions after examining an extremely small fraction of search space. Ranking is a technique to categorize & finding the best option in the market. When number of best option is available in the market, so its difficult to getting the best option is always a problem. In this paper we proposed a technique to optimize the ranking and its availability to check performance factor in order to maintained high ranking and quality of popular option in the market. We enhanced the Scheduling algorithm for ranking optimization approached, so, we used to Scheduling technique is demonstration to check

other factor which affect to ranking of products, we are finding research to get factor detail which to improve the ranking of product. This interactive technique that addresses the limitations of existing methods and is motivated by acomprehensive analysis of requirements of multi-attribute rankings considering various domains

Ranking based technique are also used to rank the data and enhanced the process of Big-Data processing. A ranking based scheduling technique is presented in this paper. Which enhances the performance of Big Data processing. Two types of ranking models called rank the query over the individual document and rank the query over the entire set of related document. In first type of ranking mechanism two models vector space model and probabilistic model are used to rank a query against individual document. To rank query against entire set of related document a set oriented model is used. That model ranks the query according to the entire set of related document.

II RELATED WORK

Rank Explorer, C. Shi, W. Cui, S. Liu, P. Xu, W. Chen, and H. Qu [1], presents a new ranking technique to reveal the ranking changes in the data. That technique contains four steps, 1) segmentation method to partition the curve into various ranking categories. 2) An enhanced Theme-River view to which contains colour bars to show the aggregation values related to each ranking category. 3) A new curve to show the degree of ranking changes. 4) Rich user interaction to support interactive changes. That method applied over real time data to provide an efficient way to show ranking changes. That technique monitor the changes over time and provide summarization for the other values.

Scheduling techniques in hadoop, Bhavsar Nikhil, Bhavsar Riddhikesh, Patil Balu, Tad Mukesh [13], a scheduler is a software which provides scheduling mechanism in Big Data data scenarios to provide an enhanced technique to process that data. There are scheduling techniques called FIFO, Fair scheduling, capacity scheduling are used to provide better performance to the user. But these techniques are not efficient to provide better performance to process Big Data. To store Big Data HDFS (Hadoop Data File System)is used because traditional data storage are not suitable store that data.

Big Data processing, Harshawardhan S. Bhosale, Devendra P. Gadekar[14], a review over the techniques used to process data over Hadoop platform is presented. Scheduling mechanisms are used to provide an enhanced mechanism processing for the Big Data. Technique like FIFO, Fair scheduling, PRISM, Capacity based scheduling etc. are used to provide scheduling mechanism for processing Big Data.

III PROPOSED METHODOLOGY

As per our observation about the previous technique and their disadvantage in different terms and scenarios. Our work present a new approach which is productive and consumes high value and thus computational better result over the large number of available dataset.

Our work propose a new algorithm Heuristic Based prediction model which utilize a new logistic normal distribution technique, which give a relation between the topics and also provide a flexible environment for the complete process and thus it generate a better prediction model for data transmission.

The proposed algorithm is described below:

1. Loading of all the available data & packets from the created given message which are participating for the communication.
2. Loading the complete node dictionary pair from the dataset.
3. Perform the particular algorithm as per selected by the user for further execution such as existing or proposed
4. Perform node down operation and matching operation if any single match is obtained and conclude that further using model for the data shifting either it is working or not.
5. Perform model and match operation if atleast 2 or more dictionary match is performed by the system.
6. Ontaining parameter wise data for the history model.
7. Observing the values and thus it effect accuracy and efficiency for the complete scenario.
8. Exit.

Algorithm Pseudo Code:

Proposed Algorithm:

Input: Node data Qi,

Output: algorithm process, Metadata, node values.

Steps:

```
Active either PRISM or Heuristic
While(true) do{
 Node distribution{p1,p2…..pN};
  dictionaryRequest();
  If(scorematching()==1)
{
Recognition();
Perform Heuristic model;
Compute the prediction values;
{
Result computation;
}
Set status=finish and exit;
} If(scorematching()>=2)
{
Re-distribution;
```


{
Computing parameter upon distribution;
}
Set status=finish and exit;
}

IV Result Analysis

As per the observed result and experiment setup, technique is implemented. The proposed and existing technique is performed with the above post which are data packet & distribution among the multiple node matching which in result given by the scheduling algorithm performed with the system and following output results were monitored:

In the table present below is a statistical comparison of the values which are retrieved as time taken by the different process algorithm, throughput and other parameter can be observe.

Table 1: Data distribution for different data packet.

Data Packets	Existing Technique	Proposed Technique
1024	3374ms	3889ms
2048	4098ms	4128ms
3072	5344ms	5229ms
4096	5391ms	5310ms

The above table represent the number of data values from the data and algorithm is performed.

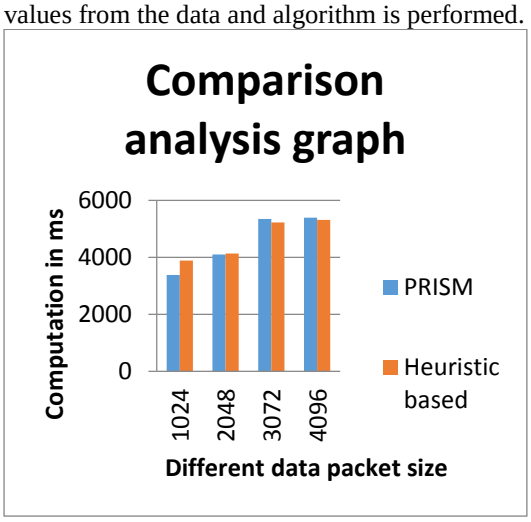

Figure 4.1: Comparison Line graph for technique analysis

In the above graph drawn x axis as data from which post were extracted for the query processing for specified dataset and line graph is printed using the chart library provided by the Microsoft and further analysis can easility performed thus the Heuristic based approach outperform the best.

V CONCLUSION

Big data termed as very large and complex data. Processing that data using traditional techniques generating huge overhead for the user. An enhanced technique is required to process that data. A ranking based scheduling technique which schedule query on the basis of ranking provided by the various users. A comparison analysis of results is presented in result and analysis section which shows, proposed technique provides an efficient mechanism to process Big Data in multiple nodes.

REFERENCES

[1] C. Shi, W. Cui, S. Liu, P. Xu, W. Chen, and H. Qu."RankExplorer: visualization of ranking changes in large time series data" IEEE, 2012.

[2] Fabrizio Angiulli, Senior Member, IEEE, Stefano Basta, Stefano Lodi, and Claudio Sartori "Distributed Strategies for Mining Outliersin Large Data Sets" IEEE, July 2013.

[3] M. Ward, G. Grinstein, and D. A. Keim. Interactive Data Visualization: Foundations, Techniques, and Application. A.K. Peters, 2010.

[4] P. Kidwell, G. Lebanon, and W. S. Cleveland. Visualizing incomplete and partially ranked data. IEEE, 2008.

[5] L. Byron and M. Wattenberg. Stacked graphs - geometry & aesthetics. IEEE, 2008.

[6] P. Sawant and C. G. Healey. Visualizing multidimensional query results using animation. In Electronic Imaging, 2008.

[7] Madria, Sanjay Kumar; 'Web Mining: A Bird's Eye View;" http://mandolin.cais.ntu.edu.sg/wise2002/slides .shtml;WISE 2002, Singapore.

[8] Baeza-Yates, Ricardo, Davis, Emilio "Web page ranking using link attributes," Proceedings of the 13th internationalWorld Wide Web conference on Alternate track papers & posters, May 2004.

[9] Xing, W.; Ghorbani, A.; "Weighted PageRank algorithm;" Proceedings of the Second Annual Conference on Communication Networks and Services Research, May 2004.

[10] A. Frome, Y. Singer, F. Sha, and J. Malik. Learning globally consistent local distance functions for shape-based image retrieval and classification. In ICCV, 2007.

[11] A. Frome, Y. Singer, and J. Malik. Image retrieval and recognition using local distance functions. In NIPS, 2006.

[12] P.-M. Cheung and J. T. Kwok. A regularization framework for multiple-instance learning. In ICML, 2006.

[13] Bhavsar Nikhil, Bhavsar Riddhikesh,Patil Balu,Tad Mukesh "A Survey On Scheduling In Hadoop For Bigdata Processing" Mjert, 2015.

[14] Harshawardhan S. Bhosale, Devendra P. Gadekar "Big Data Processing Using Hadoop: Survey on Scheduling" IJSR, 2012.

