DOI: http://dx.doi.org/10.26483/ijarcs.v8i8.4773

ISSN No. 0976-5697

Volume 8, No. 8, September-October 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

OPTIMIZED FEATURE SELECTION BASED PREDICTIVE ROUND ROBIN SCHEDULING (OFS-PRRS)

N. Arunadevi
Part-Time Research Scholar
Dept of Computer Science
Periyar University
Salem, Tamilnadu

Dr.Vidyaa Thulasiraman
Assistant Professor
Dept Of Computer Science
Govt Arts & Science College for Women
Bargur, Tamilnadu

Abstract: Optimized Feature Selection based Predictive Round Robin Scheduling (OFS-PRRS) Technique for stream data in big data analytics with higher prediction accuracy and lesser scheduling time. In OFS-PRRS technique, Least Absolute Shrinkage and Selection Operator (LASSO) function is used for feature selection. LASSO function in big data analytics is used based on assumption of linear dependency between input features and output value.

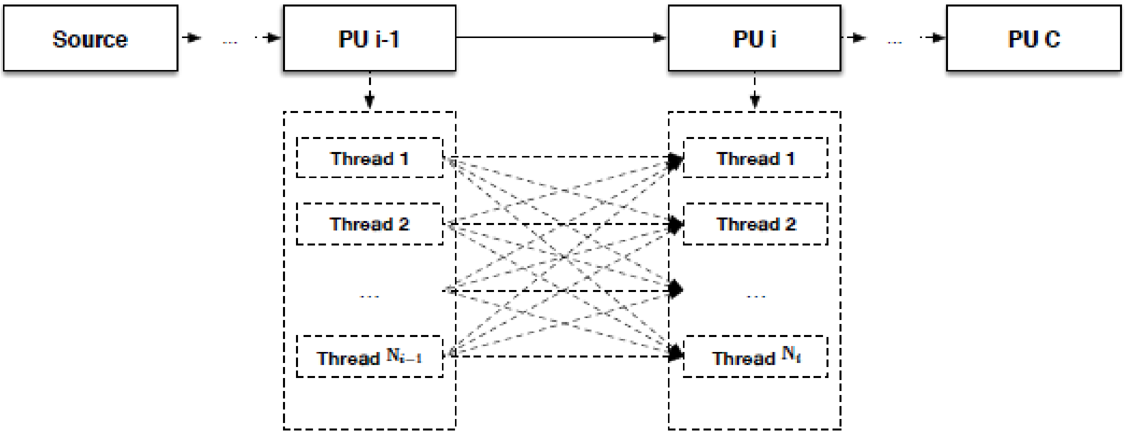
Keywords: Optimized Feature,OFS-PRRS, (LASSO),

1. INTRODUCTION

Big data is large, fast changing and dispersed beyond the ability of available hardware and software facilities for acquisition, access, analytics, and/or application in particular time and space. Big data is a term for representing large amount of data in different formats that are not handled by traditional databases. Big data is commonly used for business operations in today's competitive scenario. Big data stream computing depends on programs that compute continuous data streams. Big data analytics is the process of gathering, organizing and examining large sets of data to find out patterns or useful information. Predictive resource scheduling is used to improve the performance by leveraging big data analytics. Many research works has been designed in big data analytics of predictive scheduling for stream data. But, big data analytics is still a challenging and time demanding task for stream data predictive scheduling

2. PROBLEM DEFINITION

A topology-aware method is introduced in [1] to predict the average tuple processing time of application for given scheduling solution with topology of graph and runtime statistics. An effective algorithm is designed to allocate threads for machines based on the prediction results. Highly-regarded distributed stream data processing platform, Storm are used with three applications, namely word count, log stream processing and continuous query. However, the prediction accuracy remained unaddressed.



A new predictive scheduling framework in [2] allows fast and distributed stream data processing with the features of topology aware performance prediction and predictive scheduling. Topology-aware method is designed to predict the average tuple processing time for scheduling solution. The scheduling is carried out based on the topology of application graph and runtime statistics. Though the processing time is reduced, the computational complexity is not reduced.

A bulk data movement framework called LADS is introduced in [3] between PFS that use CCI interface for communication. LADS employ the physical view of files than logical view. Traditional file transfer tools use the logical view of files irrespective of the process where the primary objects are distributed in PFS. LADS recognize the physical layout of files. In LADS, files comprise many data objects and set of storage targets. LADS allocate all reads and writes to fundamental object size in PFS. *LADS* can avoid congested storage elements within the shared storage resource, improving I/O bandwidth, and data transfer rates across the high speed networks. But, the error rate is not reduced using bulk data movement framework.

Parallel processing structures are designed in [4] with higher processing capacity and process of parallel devices are scheduled to increase the efficiency. A dynamic assignment scheduling algorithm for big data stream processing in mobile Internet services is introduced and stream query graph calculates the weight of every edge. The edge with minimum weight is chosen to send the tuples. The system context switching is minimized by increasing number of tuples sent every time. Though the efficiency is increased, the prediction accuracy is not improved.

Optimized Feature Selection based Predictive Round Robin Scheduling (OFS-PRRS) Technique for stream data in big data analytics with higher prediction accuracy and lesser scheduling time. In OFS-PRRS technique, Least Absolute Shrinkage and Selection Operator (LASSO) function is used for feature selection. LASSO function in big data analytics is used based on the assumption of linear dependency between input features and output value. Linear Regression Prediction model is used for predicting the processing time of application based on historical data for scheduling. After predicting the future outcomes, Round Robin Scheduling is carried out in big data analytics to schedule the data with minimal scheduling time. Time quanta in Round Robin Scheduling are assigned to each process in equal portions and in circular order for obtaining predicted results without priority. Experimental evaluation is carried out on factors such as prediction accuracy, error rate and scheduling time with respect to size of big data. Round-robin algorithm is a pre-emptive algorithm as the scheduler forces the process out of the CPU once the time quota expires.

3. ACKNOWLEDGMENT

Firstly, I would like to express my sincere gratitude to my advisor Asst Prof. ***Dr. Vidyaa Thulasiraman*** *Dept of Computer Science, Govt Arts & Science college for Women Bargur* for the continuous support of my Ph.D study and related research, for her motivation, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

REFERENCES

[1] Teng Li, Jian Tang and Jielong Xu, “Performance Modeling and Predictive Scheduling for Distributed Stream Data Processing”, IEEE Transactions on Big Data, Volume 2, Issue 4, December 2016, Pages 353 – 364

[2] Teng Li, Jian Tang and Jielong Xu, “A Predictive Scheduling Framework for Fast and Distributed Stream Data Processing”, IEEE International Conference on Big Data, 2015, Pages 1-6

[3] Youngjae Kim, Scott Atchley, Geoffroy R. Vallee, Sangkeun Lee and Galen M. Shipman “Optimizing End-to-End Big Data Transfers over Terabits Network Infrastructure”, IEEE Transactions on Parallel and Distributed Systems, Volume 28, Issue 1, 2017, Pages 188 – 201

[4] Yan Liu, Kun Wang, Yue Yu, Jin Qi, Yanfei Sun, “A dynamic assignment scheduling algorithm for big data stream processing in mobile Internet services”, Personal and Ubiquitous Computing, Springer, Volume 20, 2016, Pages 373–383

[5] Dawei Sun, Guangyan Zhang, Songlin Yang, Weimin Zheng, Samee U. Khan, Keqin Li, “Re-Stream: Real-time and energy-efficient resource scheduling in big data stream computing environments”, Information Sciences, Elsevier, Volume 319, 20 October 2015, Pages 92–112

[6] Chuting Yao, Chenyang Yang and Zixiang Xiong, “Energy-saving Predictive Resource Planning and Allocation”, IEEE Transactions on Communications, Volume 64, Issue 12, December 2016, Pages 5078 – 5095

[7] Karim Kanoun, Cem Tekin, David Atienza, and Mihaela van der Schaar, “Big-Data Streaming Applications Scheduling based on Staged Multi-armed Bandits”, IEEE Transactions on Computers, Volume 65, Issue 12, December 2016, Pages 3591 – 3605

[8] Lorenz Fischer and Abraham Bernstein “Workload Scheduling in Distributed Stream Processors using Graph Partitioning”, IEEE International Conference on Big Data, 2015, Pages 124 – 133

[9] Daniel Millot and Christian Parrot “Optimization of the Processing of Data Streams on Roughly Characterized Distributed Resources”, IEEE Transactions on Parallel and Distributed Systems, Volume 27, Issue 5, May 2016, Pages 1415 – 1429

[10] Alina Sîrbu and Ozalp Babaoglu “Towards operator-less data centers through data-driven, predictive, proactive autonomics” Cluster Computing, Springer, Volume 19, Issue 2, June 2016, Pages 865–878