IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-11 E-ISSN: 2347-2693

An Investigation Study on QoS and Traffic Aware Job Scheduling Techniques with Big Data

C.R. Durga devi[1*], R. Manicka Chezian [2]

[1*]Department of Computer Science, NGM College, Pollachi, India
[2]Department of Computer Science, NGM College, Pollachi, India

**Corresponding Author: deviswe@gmail.com*

Available online at: www.ijcseonline.org

Received: 10/Oct/2017, Revised: 29/Oct/2017, Accepted: 16/Nov/2017, Published: 30/Nov/2017

Abstract— Big Data Analytics (BDA) applications are new software application for processing large amount of data to collect the hidden value. Big data is defined as a datasets whose size is beyond the capability of usual relational databases to collect, direct and handle the data with lesser latency. Most of the recent research works aimed to reduce the traffic and workload for convalescing the quality of services in big data. In recent times, many research works are carried out for getting better the performance of regression and classification process during the data access from the big data. However, the job completion time and memory space complexity remained challenging issue. Our main objective is to reduce the space complexity and time complexity during the data accessing from big data. In order to reduce the job completion time and memory space consumption, many existing techniques are reviewed. The key objective of the research is to increase performance of traffic aware job scheduling techniques with minimal space and time complexity. In this paper, review of various existing job scheduling techniques is carried out. The study and analysis about the performance of three existing techniques in terms of their space and time complexity is measured as the number of user requests increases and a comparison of the results between these techniques is carried out. Limitations of existing techniques are also discussed.

Keywords— Big Data Analytics, Relational databases, Quality of services, Regression, Classification

I. INTRODUCTION

Big data analytics has increased the attention from both industry and academic due to its large benefits in cost reduction and decision making. Big data analytics is an essential tool for changing the science, engineering, medicine, healthcare, finance and business itself. Big data analytics workload is important in modern data center to classify their workloads and to know their behaviours for improving the recital of data center. Big data analytics workload takes account of the business intelligence, machine learning and bio-informatics.

This paper is organized as follows, Section II portrays the review on different traffic aware job scheduling techniques with big data, Section III portrays the study and analysis of the existing job scheduling techniques, Section IV describes the possible comparison between them, discussion on limitations of existing techniques are studied and Section V concludes the paper. The key objective of the research is to identify the performance of Optimization Framework, Genetic Algorithm-based Job Scheduling Model and Global Architecture technique which were taken under study in terms of their memory space, job completion time and efficiency. Based on this study, the limitations are identified and it leads to development of new technique with minimal space and time complexity.

II. RELATED WORK

The inter-DC traffic created by MapReduce jobs on geo-distributed big data were minimized with predicted job completion time. The predictable job completion time were guaranteed by chance-constrained optimization technique in [1]. The designed technique finished the MapReduce job within predefined job completion time and higher probability. Though the job completion time was reduced, the traffic occurrence rate was not minimized. A genetic algorithm-based job scheduling model was introduced for big data analytics applications increased the efficiency, feasibility and accuracy in [2]. Though the job scheduling accuracy was improved, job completion time remained unaddressed.

A global architecture was designed for Quality of Service (QoS) based scheduling in big data application to the distributed cloud datacenter in [3]. The global architecture represented the complete datacenter resources in any order and performed new incoming jobs in predefined virtual clusters with QoS needs. For identifying the features of big

data analytics workload, correlation analysis was carried out to identify the factors that vary the cycles per instruction (CPI). But, the scheduling process was not carried out to reduce the workload through the correlation analysis in [4]. The big data analytics workloads share characteristics in many classes from conventional workloads and scale-out services. A new concept of DFS-integrated DBMS was introduced where DBMS was combined with distributed file system (DFS) in [5]. For processing the big data analytics in parallel, MapReduce framework called PARADISE were introduced on DFS-integrated DBMS.

In PARADISE, job splitting method splitted job depending on predicate in integrated storage system. But, the job completion time was not reduced during PARADISE framework. A novel shuffle data transfer approach addressed the issues like excessive auxiliary memory utilization and data shuffling by dynamically adapting the prefetching to computation in [6]. A new strategy in Spark was introduced called in-memory data analytics framework. But, the classification accuracy was not improved using Spark strategy. A parallel array operator depending on particular form of matrix multiplication calculated the comprehensive data summarization matrix in [7]. The matrix allowed iterative algorithms to function in main memory through reducing the number of times the dataset scanned and through minimizing the number of CPU operation. But, the workload was not reduced by parallel array operator. System-level stability evaluation model was introduced for Energy Internet with energy function to search small disturbance stability region (SDSR) in [8]. SDSR were obtained through calculating the operational data threshold of distributed generations (DGs). Though energy consumption was reduced, the traffic occurrence rate was not reduced using system-level stability evaluation model. A distributed cloud-computing framework was introduced with Big Data approach in which storage and computing resources were used to gather and process traffic from large-scale network with minimal time consumption in [9]. Though the time complexity was reduced, traffic occurrence rate was not reduced in effective manner. Hadoop is an open source framework that process large quantity of data in efficient manner. Job scheduling is an essential factor for attaining better results in big data processing. Hadoop Distributed File System (HDFS), Hadoop MapReduce and various parameters were highlighted that increased the performance of job scheduling algorithms in big data like Job Tracker, Task Tracker, Name Node, Data Node, etc in [10].

Hadoop is a quickly budding ecosystem of components based on Google's MapReduce algorithm and file system work for implementing MapReduce algorithms in a scalable fashion in [11]. In multi-cluster environment, Grey Wolf Optimization algorithm was applied in [12].

III. METHODOLOGY

Qos and traffic aware job scheduling techniques with big data

Big data analytics attracted large attention from both industry and academic due to its great advantages in cost reduction and better decision making. MapReduce is programming model for big data processing on large clusters comprising of thousand machines. It includes two types of tasks, namely map tasks and reduce tasks. The input data are partitioned into independent chunks that are processed by map tasks in parallel. The generated key-value pairs are shuffled to minimize the tasks for producing the final results. MapReduce are implemented by many systems that are organized in single-cluster situation for big data analytics. For processing the data stored in multiple geodistributed clusters, new challenges are imposed by geo-distributed environment where the inter-cluster network connection is bottleneck.

Traffic-aware Geo-distributed Big Data Analytics with Predictable Job Completion Time

Geo-distributed data analytics aggregates the data stored in multiple data centers into single data center and then process using Hadoop or Spark that employed MapReduce model. Inter-DC traffic of MapReduce jobs targeting geo-distributed big data was minimized. A new optimization framework was introduced through considering input data movement and task placement. Input data at data center are loaded by map tasks at additional data centers when remote data loading minimizes the total inter-DC traffic. For guaranteeing the predictable job completion time rather than OI-ratio estimation, chance-constrained optimization framework required small amount of information regarding the distribution of OI-ratio. An efficient algorithm is introduced through addressing the demands. A joint optimization of data movement and task assignment addressed the issue as nonlinear.

A linearization technique is used to replace the nonlinear limitations with the linear ones. The chance constraint to achieve the predicted job completion time is not solved by convex optimization technique. An approximation approach is designed for addressing the limitations such that the solution of new formulation is feasible to the original problem.

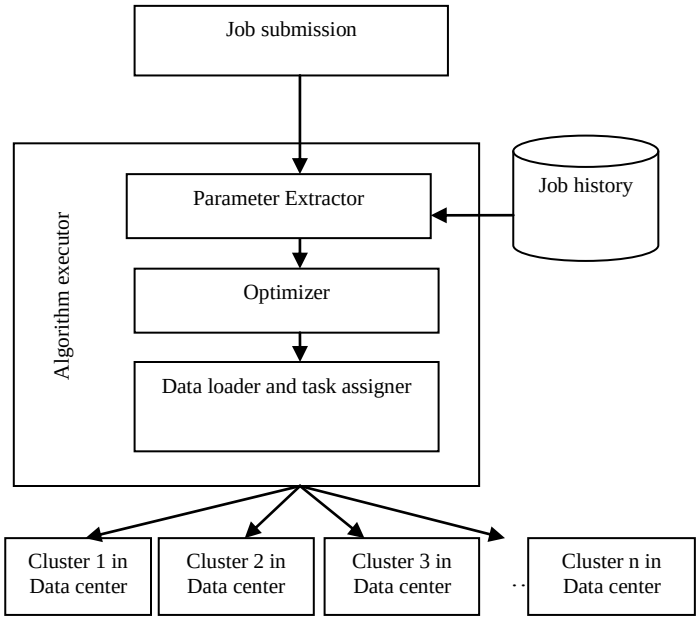

Figure 1 Optimization Framework Design

When the MapReduce job is submitted, the algorithm executor is accommodated in engines. From figure 1, the designed algorithm extracts the system parameters from job needs imposed by users and job history. The optimizer runs algorithm to choose the input data loading and task placement that are used by final module. Parameter extractor calculates the parameters required by optimizer for optimal data and task placement. It computes the bandwidth between the clusters through sending the probe packets. OI-ratio of map tasks are computed by examining the historical execution records of similar jobs.

After collecting the parameters, the optimizer identifies the input data movement and task placement for reducing the job completion time. As the inter-cluster network is inadequate resource shared by many applications, the algorithm reduces total inter-cluster traffic acquired by MapReduce job. The designed algorithm is demanding one due to joint consideration of data fetching scheme, task placement and uncertainty of OI ratio. Then, data loader retrieves input data according to the results returned by the optimizer. In addition, the task assigner starts map task for every input split and reduces tasks whose placement is identified by optimizer. A map task is placed at same machine storing target input split though the machine has enough resources. If not, the task is placed at other machines in same rack. When the map task failed to be accommodated in same rack, it is placed at additional racks in similar cluster where there are sufficient computational resources.

A genetic algorithm-based job scheduling model for big data analytics

Genetic Algorithm based decision-making is introduced for job scheduling. The estimation module is employed for clusters and jobs. The job execution information like time and cost array are gathered which explains time and cost taken by every job run on all clusters. The time and cost information is employed in framework where GAs provides an optimized solutions for job scheduling scheme. But, dissimilar processing jobs failed to have similar performance with same cluster configuration because of job characteristics. In addition, one job failed to obtain the reliable result with many clusters due to the cluster or Hadoop features. While allocating the multiple jobs to process data in one data center, there are many optional cluster configuration circumstances for every job. The selection of best job scheduling scheme is another key requirement. For getting the optimized solutions in job scheduling decision, genetic algorithm is introduced to choose the solutions with minimal time consumption and computational cost. Architecture is introduced to schedule the big data application requests to suitable datacenter at coarse grained level and to effective virtual cluster at fine grained level b⋯ ⁻¹obal scheduler and local scheduler respectively. A ⋯ ve K-nearest neighbor (AKNN) algorithm in global scheduler identifies the suitable local datacenter depending on user location and needs. Big data request are divided into computational intensive, memory intensive and input/output intensive depending on QoS needs by naïve Bayes algorithm. Self Organizing Maps (SOM) technique in local datacenter forms virtual clusters for particular kind of big data request. SOM generates topological ordering of virtual clusters where the virtual clusters are related to each other. Every incoming big data request is assigned to its particular virtual cluster for its better performance and effective resource scheduling.

A genetic algorithm-based job scheduling model was introduced for geo-distributed data with the big data analytics applications to increase the efficiency. For executing the job scheduling model, estimation module predict the execution time and cost performance of data processing jobs by GAs depending on different cluster characteristics. The job scheduling model was introduced for improving the feasibility and accuracy.

Scheduling of Big Data Applications on Distributed Cloud based on QoS parameters

Scheduling of cloud resources according to QoS parameters is very important for big data applications. The designed framework schedules the big data applications over geographically distributed cloud datacenters. Two different schedulers are used for efficient scheduling of cloud resources. Global scheduler is used at a coarse grain level and local scheduler is used at a fine grained level.

The schedulers are responsible for choosing appropriate datacenter and cluster for big data application request.

Multiple functional and quality attributes are linked with any cloud datacenter and big data processing request from any user. When the amount of data in big data applications is large, datacenter selection decision at coarse grained level considered the parameters like physical distance, average resources and network throughput (GB/s). The choice of high datacenter network throughput is not helpful when it is not well-suited with user's network. For big data applications, three parameters are taken to choose the local datacenter, namely physical distance, network throughput and available resources. With large quantity of transferred data, physical distance and network throughput are taken in selection of local datacenter. Though physical distance and network throughput are advantageous, local datacenter without resource failed to present desired QoS needs.

A global architecture was introduced for QoS based scheduling in big data application to distributed cloud datacenter in both coarse grained level and fine grained level. At coarse grain level, suitable local datacenter was selected with network distance between the user and datacenter, network throughput and accessible resources by adaptive K nearest neighbor algorithm. In fine grained level, probability triplet (C, I, M) was predicted by naïve Bayes algorithm. The algorithm provides the probability of new application in compute intensive (C), input/output intensive (I) and memory intensive (M) types. Every datacenter is changed into pool of virtual clusters for execute particular category of jobs with exact (C, I, M) requirements by self organized maps. The global architecture represents the complete datacenter resources in any ordering and executing incoming jobs in predefined virtual clusters depending on QoS requirements.

IV. RESULTS AND DISCUSSION

Comparison of QOS and Traffic Aware Job Scheduling Techniques with Big Data & Suggestions

In order to compare QoS and traffic aware job scheduling techniques, no. of cloud user request is taken to perform the experiment. Various parameters are used for improving the performance of traffic minimization and job scheduling techniques with big data.

Space Complexity

Space complexity (SC) is defined as the amount of memory space used for storing the cloud user data in cloud server for further accessing. It is measured in terms of megabytes (MB). The space complexity formula is given by,

SC = no. of cloud user * memory space consumed for one cloud user data

When the space complexity is lesser, the method is said to be more efficient.

Table 1. Space Complexity

Number of cloud users (number)	Space Complexity (MB)		
	Optimization Framework	Genetic Algorithm-based Job Scheduling Model	Global Architecture
10	21	28	33
20	25	32	36
30	27	35	39
40	29	37	43
50	31	40	46
60	35	42	49
70	38	44	52
80	41	47	55
90	45	50	58
100	48	53	61

Table 1 describes the space complexity with respect to number of cloud user ranging from 10 to 100. Space complexity comparison takes place on existing Optimization Framework, Genetic Algorithm-based Job Scheduling Model and Global Architecture. From the table value, it is clear that the space complexity using Optimization Framework is lesser when compared to Genetic Algorithm-based Job Scheduling Model and Global Architecture. The graphical representation of space complexity is shown in figure 2.

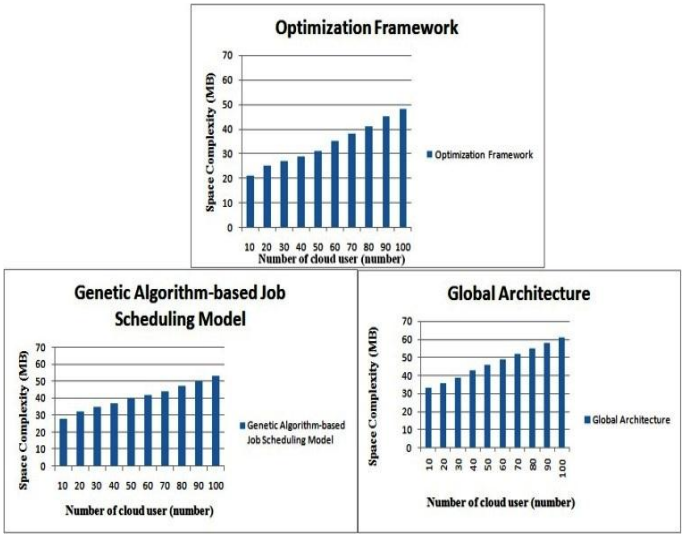

Figure 2 Measure of Space Complexity

From figure 2, space complexity based on the different number of cloud user is described. From the figure 2, Optimization Framework consumed lesser memory space than Genetic Algorithm-based Job Scheduling Model and Global Architecture. Optimization Framework Model consumed 16% lesser memory space than Genetic Algorithm-based Job Scheduling and consumed 41% lesser memory space than Global Architecture

Job Scheduling Time

Job Scheduling Time (JST) is defined as the amount of time taken for scheduling the jobs based on the number of cloud user requests. It is measured in terms of milliseconds (ms). The job scheduling time is mathematically formulated as,

JST = Ending time – Starting time for scheduling jobs

When the job scheduling time is lesser, the method is said to be more efficient.

Table 2. Job Scheduling Time

Number of cloud user requests (number)	Job Scheduling Time (ms)		
	Optimization Framework	Genetic Algorithm-based Job Scheduling Model	Global Architecture
10	45	31	57
20	49	36	60
30	53	40	64
40	57	43	67
50	59	47	70
60	63	51	73
70	67	55	76
80	71	59	80
90	75	63	84
100	78	67	87

Table 2 describes the job scheduling time with respect to number of cloud user ranging from 10 to 100. Job scheduling time comparison takes place on existing Optimization Framework, Genetic Algorithm-based Job Scheduling Model and Global Architecture. From the table value, it is clear that the job scheduling time using Genetic Algorithm-based Job Scheduling Model is lesser when compared to Optimization Framework and Global Architecture. The graphical representation of job scheduling time is described in figure 3.

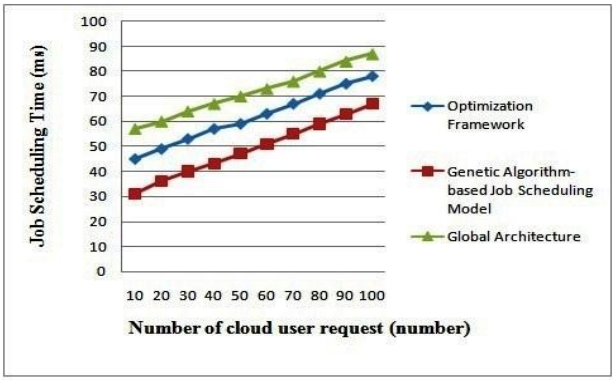

Figure 3 Measure of Job Scheduling Time

=

From figure 3, job scheduling time based on the different number of cloud user requests is described. From the figure 3, Genetic Algorithm-based Job Scheduling Model consumed lesser time consumption for job scheduling than Optimization Framework and Global Architecture. Research in Genetic Algorithm-based Job Scheduling Model consumed 20% lesser job scheduling time than Optimization Framework and consumed 31% lesser job scheduling time than Global Architecture.

Job Scheduling Efficiency (JSE)

Job scheduling efficiency is defined as the ratio of number of jobs scheduled efficiently to the total number cloud user requests. It is measured in terms of percentage. The job scheduling efficiency is mathematically formulated as,

$$JSE = \frac{\text{Number of jobs scheduled efficiently based on user requests}}{\text{Number of cloud user requests}}$$

When the job scheduling efficiency is higher, the method is said to be more efficient

Table 3 Job Scheduling Efficiency

Number of cloud user requests (number)	Job Scheduling Efficiency (%)		
	Optimization Framework	Genetic Algorithm-based Job Scheduling Model	Global Architecture
10	65.12	71.23	79.96
20	67.54	73.87	81.45
30	69.98	75.96	84.63
40	71.45	79.47	86.79
50	75.87	82.19	88.71
60	77.96	85.36	91.96
70	79.36	87.71	93.41
80	81.58	89.63	95.39
90	82.98	91.89	96.89
100	84.63	93.73	97.36

Table 3 describes the job scheduling efficiency with respect to number of cloud user ranging from 10 to 100. Job scheduling efficiency comparison takes place on existing Optimization Framework, Genetic Algorithm-based Job Scheduling Model and Global Architecture. From the table value, it is clear that the job scheduling efficiency using Global Architecture is more when compared to Optimization Framework and Genetic Algorithm-based Job Scheduling Model. The graphical representation of job scheduling efficiency is described in figure 4.

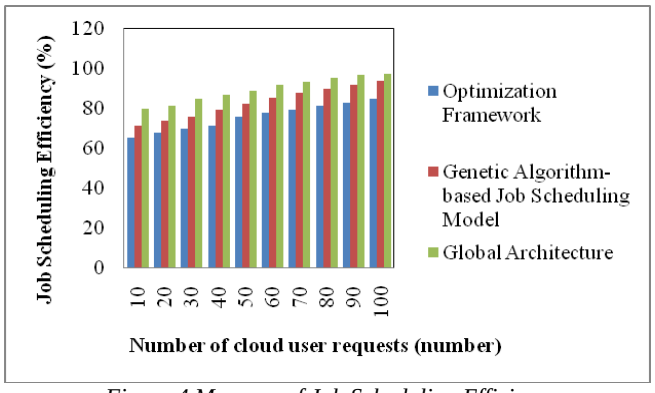

Figure 4 Measure of Job Scheduling Efficiency

From figure 4, job scheduling efficiency based on the different number of cloud user requests is described. From the figure 4, Global Architecture has higher job scheduling efficiency than Optimization Framework and Genetic Algorithm-based Job Scheduling Model. Research in Global Architecture has 19% higher job scheduling efficiency than Optimization Framework and has 8% higher job scheduling efficiency than Genetic Algorithm-based Job Scheduling Model.

Discussion on limitation of QOS and traffic aware job scheduling techniques with big data

An optimization framework minimizes the inter-DC traffic occurred by MapReduce jobs targeting on geo-distributed big data within lesser space complexity. An algorithm with linearization and relaxation techniques addresses the approximation issue by off-the-shelf solvers. By using optimization framework, MapReduce job finishes within predefined completion time. But, the integration of designed algorithm into data processing platforms was not implemented. The system implementation by optimization technique with popular data processing platforms was not carried out.

A genetic algorithm-based job scheduling model was designed for big data analytics to complete the scheduling with lesser time. For implementing the job scheduling model, an estimation module was employed to predict the clusters when performing the analytics jobs. Genetic algorithm-based job scheduling model provides efficient information estimation for geo-distributed data. Though the job completion time was lesser, estimation module was not simplified and not precise.

A global architecture is constructed for QoS based scheduling in big data application to distributed cloud datacenter with more efficiency than other methods. Every datacenter is changed into virtual clusters for executing particular category of jobs with exact (C, I, M) needs by self organized maps. The global architecture symbolizes the entire datacenter resources in predefined ordering and executing new incoming jobs in predefined virtual clusters depending on QoS requirements.

V. CONCLUSION AND FUTURE SCOPE

A comparison of different existing traffic aware scheduling techniques with big data is studied. From the study, it is observed that the existing techniques undergo heavy traffic during the data access in big data. The survival review shows that the existing optimization framework minimizes the inter-DC traffic generated by MapReduce jobs targeting on geo-distributed big data with minimal space complexity when compared to other methods under study. Energy consumption was not considered while reducing the traffic. In addition, scheduling process was carried out in effective manner using genetic algorithm-based job scheduling model with lesser job scheduling time but the memory consumption was not reduced. Global Architecture has higher job scheduling efficiency than Optimization Framework and Genetic Algorithm-based Job Scheduling Model as the number of user request increases. In this method, the load balancer at global and local scheduler level failed to balance the load effectively among data centers and virtual clusters depending on dynamic QoS requirements of big data application. In addition, time consumption and space consumption was not reduced Finally, from the result, each technique has its own merit and demerit and based on them new techniques can be devised to concentrate more on minimizing the time and space complexity during job scheduling for big data applications. The future path of QoS and traffic aware job scheduling can be carried out using machine learning techniques for increasing the scheduling efficiency and minimizing the traffic with minimal time and space complexity.

References

[1] Peng Li, Song Guo, Toshiaki Miyazaki, Xiaofei Liao, Hai Jin, Albert Y. Zomaya,Kun Wang, "*Traffic-aware Geo-distributed Big Data Analytics with Predictable Job Completion Time*", IEEE Transactions on Parallel and Distributed Systems, Volume 28, Issue 6, Pages 1785 – 1796, 2017.

[2] Qinghua Lu, Shanshan Li, Weishan Zhang and Lei Zhang, "*A Genetic Algorithm-Based Job Scheduling Model for Big Data Analytics*", EURASIP Journal on Wireless Communications and Networking, Springer, Volume 16, Issue 152, Pages 1-9, 2016

[3] Rajinder Sandhu and Sandeep K. Sood, "*Scheduling Of Big Data Applications on Distributed Cloud Based on Qos Parameters*", Cluster Computing, Springer, Volume 18, Issue 2, Pages 817–828, June 2015

[4] Zhen Jia, Jianfeng Zhan, Lei Wang, Chunjie Luo, Wanling Gao, Yi Jin, Rui Han and Lixin Zhang, "*Understanding Big Data Analytics Workloads on Modern Processors*", IEEE Transactions on Parallel and Distributed Systems, Volume 28, Issue 6, Pages 1797 – 1810, June 2017

[5] Jun-Sung Kim, Kyu-Young Whang, Hyuk-Yoon Kwon and Il-Yeol Song, "*PARADISE: Big Data Analytics using the DBMS Tightly Integrated with the Distributed File System*", World Wide Web, Springer, Volume 19, Issue 3, , Pages 299–322, May 2016

[6] Bogdan Nicolae, Carlos H. A. Costay, Claudia Misalez, Kostas Katrinis and Yoonho Park, "*Leveraging Adaptive I/O to Optimize Collective Data Shuffling Patterns for Big Data Analytics",* IEEE Transactions on Parallel and Distributed Systems, Volume 28, Issue 6, Pages 1663 – 1674, June 2017

[7] Carlos Ordonez, Yiqun Zhang and Wellington Cabrera, "*The Gamma Matrix to Summarize Dense and Sparse Data Sets for Big Data Analytics",* IEEE Transactions on Knowledge and Data Engineering, Volume 28, Issue 7, Pages 1905 – 1918 July 2016,

[8] Kun Wang, Huining Li, Yixiong Feng, and Guangdong Tian, "*Big Data Analytics for System Stability Evaluation Strategy in the Energy Internet",* IEEE Transactions on Industrial Informatics, Volume 13, Issue 4, Pages 1969 – 1978, August 2017

[9] L. U. Laboshin, A. A. Lukashin and V. S. Zaborovsky, "*The Big Data Approach to Collecting and Analyzing Traffic Data in Large Scale Networks",* Procedia Computer Science, Elsevier, Volume 103, Pages 536-542, 2017.

[10] Mohd Usama, Mengchen Liu and Min Chen, "*Job schedulers for Big data processing in Hadoop environment: Testing real-life schedulers using benchmark programs",* Digital Communications and Networks, Elsevier, Pages 1-14, August 2017.

[11] E. Sivaraman, Dr.R.Manickachezian, "*High Performance and Fault Tolerant Distributed File System for Big Data Storage and Processing Using Hadoop",* IEEE Xplore Digital Library, DOI: 10.1109/ICICA.2014.16**,** E-ISBN**:** 978-1-4799-3966-4

[12] Sapinderjit Kaur, Kirandeep Kaur, Amit.Chhabra, *"Parallel Job Scheduling Using Grey Wolf Optimization Algorithm For Heterogenous Multi-Cluster Environment",* International Journal of Computer Science and Engineering Vol.5 , Issue.10 , pp.44-53, Oct-2017

[13] S.Hemalatha, Dr.R.Manickachezian, "*Implicit Security Architecture Framework in Cloud Computing Based on Data Partitioning and Security Key Distribution*", International Journal of Emerging Technologies in Computational and Applied Sciences, pp. 76-81, ISSN: 2279-0055, Feb. 2013.

Authors Profile

C.R.Durga devi received her Bsc.Computer Technology from Coimbatore Institute of Technology,Coimbatore,India. she had her Master of Computer Applications from Bharathiar University,Coimbatore, India. she holds Mphil in Computer Science from Bharathiar University, Coimbatore, India. She has 11 years of experience in teaching. She is presently working as an Assistant Professor in NGM college, pollachi. Her research interest includes Data Mining,Big Data Analytics. Now she is pursuing her ph.d Computer Science in Dr.Mahalingam center for research and Development at NGM college, Pollachi.

Dr.R.Manickachezian received his M.Sc., Degree in Applied Science from P.S.G College of Technology, Coimbatore, India in 1987. He completed his M.S. Degree in Software Systems from Birla Institute of Technology and Science, Pilani, Rajasthan, India and Ph.d degree in Computer Science from school of Computer Science and Engineering, Bharathiar University, Coimbatore, India. He served as a faculty of maths and Computer Applications at P.S.G College Of Technology, Coimbatore from 1987 to 1989. presently, he has been working as an Associate Professor of Computer Science in NGM college (autonomous), pollachi under Bharathiar University, Coimbatore, India since 1989. He has published 150 papers in International/National Journal and Conferences. He is a recipient of many awards like best Computer Science Faculty of the year 2015, Best Research Supervisor award, Life Time Achievement award, Desha Mithra Award and best paper award. His research focuses on Network Databases, Data Mining, Distributed Computing, Data Compression, Mobile Computing, Real Time Systems and Bio-Informatics.

